

Statistiques descriptives

L'ÉCONOMIE ET LES CHIFFRES

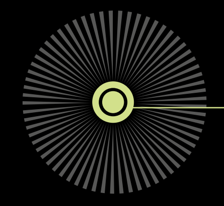
P. Bailly et C. Carrère



Les chiffres constituent une part majeure de l'information économique et sociale, en particulier les données statistiques, dont médias et gouvernants usent et abusent : elles sont un outil incontournable de connaissance de la réalité. Pour autant, les chiffres ne disent rien si on ne sait pas les faire parler.

Ainsi, la statistique suppose une bonne maîtrise de la théorie afin de produire des données cohérentes puis une analyse fine des résultats qui, seules, permettent de faire parler les chiffres.

En présentant l'essentiel des outils à connaître (les tendances centrales, de dispersion, de concentration, les ajustements, le traitement des chroniques, la présentation des indices), l'ouvrage constitue une excellente introduction aux traitements statistiques pour les étudiants de 1er cycle en économie-gestion (Licence, IUT, BTS). Il associe théorie et cas pratiques avec de nombreux exemples concrets et en contexte, afin d'éclairer la théorie par l'illustration.



PIERRE BAILLY enseigne les statistiques descriptives à la Faculté d'économie de Grenoble de l'université Grenoble Alpes.

CHRISTINE CARRÈRE est enseignante de statistiques.



Presses universitaires de Grenoble
BP 1549 – 38025 Grenoble cedex 1
ISBN 978-2-7061-2215-6 *(e-book PDF)*



Statistiques descriptives



Création de couverture : Corinne Tourrasse

5, place Robert-Schuman
BP 1549 – 38025 Grenoble cedex 1
pug@pug.fr / www.pug.fr

ISBN 978-2-7061-2215-6 *(e-book PDF)*
L'ouvrage papier est paru sous la référence ISBN 978-2-7061-2214-9.

Pierre Bailly
Christine Carrère

Statistiques descriptives

L'économie et les chiffres

Presses universitaires de Grenoble

Dans la même collection

Droit

O. Soria, *Droit de l'environnement industriel*, 2013
M. Pérès, *Droit et responsabilité en montagne. Jurisprudence relative aux activités sportives et touristiques en montagne*, 2006
D. Mallet, P. Balme, P. Richard (dir.), *Réglementation et management des universités françaises*, 2005
P. Pedrot (dir.), *Génétique, biomédecine et société*, 2005
F. Servoin, *Droit administratif de l'économie*, 2001

Économie

R. Colliat et Y. Échinard (dir.), *Quelle fiscalité pour le XXIe siècle ? Contributions au débat,* 2014
P. Bailly, C. Carrère, *Statistiques descriptives. Cours*, 2e édition, 2007
P. Bailly, C. Carrère, *Statistiques descriptives. Exercices avec corrigés*, 2e édition, 2007
H. Drouvot, *Le Made in Brésil. L'industrie brésilienne face à la mondialisation*, 2005
C. Perret (dir.), *Perspectives de développement pour la Nouvelle-Calédonie*, 2002
M. Lejeune, *Traitements des fichiers d'enquêtes. Redressements, injections de réponses, fusions*, 2001
A. Samuelson, *Les grands courants de la pensée économique*, 5e édition, 1997

Gestion

P. Balme, J.-R. Cytermann, M. Dellacasagrande, J.-L. Reffet, P. Richard, D. Verhaeghe, *L'université française : une nouvelle autonomie, un nouveau management*, 2012
B. Derrouch, *Gestion comptable et financière de l'entreprise*, 2005

Sciences politiques

Y. Deloye, O. Ihl, A. Joignant (dir.), *Gouverner par la science : perspectives comparées*, 2013
G. Gourgues, *Les politiques de démocratie participative*, 2013
M. Hollard, G. Saez (dir.), *Politique, science et action publique. La référence à Pierre Mendès France et les débats actuels*, 2010
C. Bidégaray, S. Cadiou et C. Pina, *L'élu local aujourd'hui*, 2009
M. Chauchat, *Vers un développement citoyen. Perspectives d'émancipation pour la Nouvelle-Calédonie*, 2006
J.-L. Chabot, *Aux origines intellectuelles de l'Union européenne. L'idée d'Europe unie de 1919 à 1939*, 2005

Psychologie

E. Grebot, *Repères en psychopathologie*, 2002
E. Grebot et I. Orgiazzi Billon-Galland, *Les bases de la psychopathologie – Éléments historiques, notionnels et théoriques*, 2001

Sciences

G. Dhont, B. Zhilinski, *Symétrie dans la nature*, 2011

Sociologie

R. Levy, C. Soldano et P. Cuntigh (dir.), *L'université et ses territoires. Dynamismes des villes moyennes et particularités de sites*, 2015
A. Baron, *Innover dans les politiques sociales. Pratiques du changement*, 2013
A. Baron, *Dynamiques territoriales de l'action sociale et médico-sociale*, 2010
F. Moutet, *La Féminisation des effectifs chirurgicaux*, 2010

Introduction

Les observations numériques, en particulier celles d'ordre statistique, participent massivement de l'information économique et sociale dont les médias et les gouvernants usent et abusent. Les données statistiques sont le complément indispensable pour les démonstrations économiques, elles justifient d'accepter ou de rejeter telle ou telle analyse. Elles constituent un outil incontournable de connaissance de la réalité économique[1], en particulier parce qu'elles permettent de quantifier des phénomènes et donc d'estimer leur importance, ce que n'autorise pas une approche purement qualitative. Cependant, la statistique est tout autant science de la classification que de la quantification. En ce sens, elle ne se réduit pas à la production de nombres.

La maîtrise des méthodes du calcul statistique ne suffit pas pour produire des résultats intéressants ; la qualité des résultats est fonction de la qualité des données à l'entrée souvent exprimée par la formule anglaise *garbage in, garbage out*. L'évaluation de la qualité des données utilisées, associée à une bonne connaissance des concepts et des notions est primordiale pour apprécier la pertinence des résultats obtenus et en proposer un commentaire. Les chiffres fournissent la mesure de variables économiques, ils n'expliquent ni comment ni pourquoi elles prennent telle ou telle valeur, encore moins les conséquences économiques de la grandeur de celles-ci. La signification des chiffres est une des conditions premières de la compréhension de la situation économique.

Le champ de cet ouvrage se concentre sur la statistique dont l'objet est de fournir une description ; ni les lois de probabilités ni les statistiques inférentielles et de la décision ne seront traitées. L'inférence statistique, la prévision et l'estimation qui constituent la suite logique des domaines examinés ne seront pas traitées dans cet ouvrage.

1. Cette réalité n'est pas indépendante des moyens utilisés pour l'appréhender.

Qu'est-ce que la statistique ?

Le terme même de statistique est largement polysémique, quelques explications sont nécessaires pour préciser les différentes notions. Comme pour chaque champ de la connaissance, il existe un vocabulaire technique spécifique dont la compréhension est indispensable pour apprécier la pertinence des données utilisées. Les statistiques sont le produit d'une démarche, d'une construction théorique. Elles ne peuvent exprimer toute la complexité de la réalité économique qui elle-même ne peut être réduite aux aspects quantitatifs. Contrairement à la conception spontanée et illusoire qu'il suffit de faire une enquête pour obtenir les bonnes informations qui seraient déjà disponibles prêtes à se révéler à l'enquêteur impartial, le moindre questionnaire demande, pour apporter des données utilisables, une réflexion sur les dispositifs pertinents à mettre en œuvre en vue de cet objectif. Le modèle paradigmatique de cette conception reste le recensement, une opération difficile et coûteuse, qui fournit de précieuses et fiables observations avec, néanmoins, des ordres des marges d'erreur estimées.

Cet ouvrage portera sur la production d'informations statistiques dans le champ de l'économie et du social, il laisse donc de côté la statistique mathématique. La science statistique ne se réduit pas à une simple technique d'analyse et de présentation d'informations économiques et sociales quantifiées : un calcul, aussi simple soit-il, n'a de sens que par rapport à des données – validité, fiabilité, précision – et en vue d'un objectif.

Les données n'existent pas dans tous les domaines car la production de statistiques répond à une ou des demandes, plus ou moins vagues, qui doivent être formalisées par l'institution statistique et les statisticiens. Les choix des données à analyser et les orientations d'investigation sont arbitrés au niveau des décideurs, que ce soient des institutions publiques ou privées. S'ajoutent à ces contraintes « politiques » des contraintes budgétaires qui orientent la production de statistiques du système public comme des acteurs privés. Enfin les conditions de mobilisation, de recherche et de traitement des données dépendent de l'existence et de la qualité des organismes assurant ces missions. La neutralité et l'objectivité des statistiques sont néanmoins assurées par l'implication des professionnels de la production des données. La production statistique a pour premier objectif de fournir des observations puis dans un second temps d'assurer le traitement de celles-ci. Le second aspect est souvent le seul développé dans les manuels, ce qui tend à réduire les statistiques à un ensemble de techniques mathématiques plus ou moins complexes sans aucune interrogation préalable sur les observations faisant l'objet des calculs.

Quelques définitions

La statistique est un ensemble de principes et de méthodes scientifiques pour recueillir, classer, synthétiser et communiquer des données numériques en vue de leur utilisation pour en tirer des conclusions et prendre des décisions. La diversité des usages du mot « statistique » reflète la double nature des pratiques sociales qui lui sont associées. À l'activité administrative d'élaboration des données se combine la réflexion scientifique mathématique. Autrefois imbriquées, les deux significations s'autonomisent au début du XIXe siècle. Le terme de statistique est riche de significations, au singulier c'est un ensemble de techniques mathématiques de traitement des données numériques. La statistique renvoie à une méthode scientifique, une branche des mathématiques[2] dont les principes découlent de la théorie des probabilités et qui a pour objet le groupement méthodique ainsi que l'étude des séries de faits ou de données numériques. Au pluriel, les statistiques sont synonymes de nombres, de données, d'informations numériques, elles indiquent une pluralité de phénomènes à travers les nombres attachés à l'appréhension de ceux-ci. Avec un article indéfini : une statistique est une série de nombres parmi d'autres séries possibles. De plus, le terme de statistique désigne fréquemment une série numérique, nous emploierons plus volontiers l'expression de distribution statistique.

La connaissance statistique est le rapport entre un besoin d'information et les moyens disponibles pour les produire. Elle se situe à la jonction d'une démarche théorique et d'une démarche empirique, un lieu assez peu confortable. La mesure dépend de conventions portant sur la définition de l'objet et les procédures de codages souvent d'origines administratives. C'est en particulier le cas des nomenclatures qui finissent parfois comme les catégories socioprofessionnelles par devenir hégémoniques dans l'appréhension de la réalité à décrire. Elle résulte également de l'application de concepts théoriques produits par des travaux scientifiques. Ces deux sources peuvent parfois fournir des résultats divergents. L'appréciation du chômage en est un exemple emblématique avec les deux estimations concurrentes, d'une part le nombre de demandeurs d'emploi en fin de mois, DEFM, mesurés par Pôle emploi dans une logique de gestion, et d'autre part le chômage au sens du Bureau international du travail (BIT), résultat d'enquêtes de l'INSEE pour des comparaisons internationales.

2. Ce n'est qu'avec l'autonomisation de ce champ de recherche, crée par Galton et Pearson et introduite en France par Lucien March, au début du XXe siècle que ce sens est admis. A. Girard, « La recherche en histoire de la statistique », *Courrier des Statistiques*, n° 32, octobre 1984.

Un calcul statistique n'a de sens que pour l'étude de populations importantes et que s'il permet d'obtenir un gain d'information ou de signification par rapport à une analyse immédiate des données. Les opérations de calcul réduisent notablement le volume des informations initiales. Les méthodes de la statistique descriptive ambitionnent de condenser un ensemble de données numériques en quelques indicateurs significatifs.
Le traitement statistique est aujourd'hui effectué automatiquement à l'aide d'ordinateurs. Le traitement de grandes quantités de données s'accomplit aisément et dans des délais très brefs, quasi instantanés. La compréhension, l'effort d'assimilation et d'appropriation des méthodes statistiques constituent, néanmoins, un préalable à une utilisation efficace et pertinente de ces outils. C'est pourquoi nous ne présentons aucun traitement sur ordinateur.

Une courte histoire

Si le mot statistique est relativement récent puisqu'il semble avoir été introduit en Allemagne au XVIIe siècle, la pratique des statistiques est par contre ancienne. Les grands empires antiques centralisateurs et unificateurs ont été, en raison de leur nature, confrontés à la nécessité du dénombrement des hommes et des biens pour les grands travaux et la levée des armées que ce soit en Mésopotamie, dans l'Égypte ancienne, en Chine antique comme dans l'Empire indien. Il s'agissait selon les époques de connaître la population pour la répartir sur le territoire, distribuer les terres, établir les rôles d'imposition, des corvées, des conscriptions militaires… Les dénombrements dans ces civilisations ont certes des significations fiscales ou militaires, mais aussi une portée magique ou religieuse, les unes et les autres fortement imbriquées.
En Grèce, les dénombrements distinguent les hommes libres, les métèques et les esclaves. La réflexion porte sur le nombre idéal de citoyens que doit comporter une cité et sur les moyens de le maintenir. À Rome, les recensements sont périodiques tous les cinq ans puis tous les dix ans. Ils permettaient de déterminer qui était citoyen romain, et de classer les citoyens d'après leurs revenus et de lever l'impôt. Ils étaient aussi le moyen de définir la place de chacun dans l'organisation politico-administrative militaire de la cité.
En France, le raffermissement du pouvoir royal avec les Carolingiens amène à un renouveau des inventaires (les capitulaires) de tous leurs biens (hommes, habitations, céréales, bétail). En Angleterre au XIe siècle, on procède à un ensemble de relevés afin de produire un cadastre. On recense les noms de lieux, de leurs tenanciers, le nombre d'occupants de chaque demeure, celui

des serfs, des hommes libres et de l'étendue des terres. Ce recensement ne concerne ni le clergé, ni les femmes, ni les enfants, ni les pauvres. Dans ces cas, l'unité sociale de base est le feu non la personne.

Le XVIe siècle est celui où les États se centralisent et s'unifient par la recherche d'une cohérence interne. C'est l'époque où règne le mercantilisme et où les auteurs cherchent à mesurer la richesse du Prince. J. Bodin, mercantiliste français, expose les avantages d'une meilleure connaissance de la population du royaume que ce soit pour la guerre ou pour la fiscalité. Le dénombrement des biens s'avère indispensable pour que la charge fiscale de chacun soit équitable, afin d'éviter des troubles et guerres civiles. La théorie mercantiliste développe la thèse selon laquelle l'État accroît sa force en favorisant l'enrichissement des citoyens. L'augmentation de la population, qui permet de maintenir de bas salaires, est, elle aussi, source de richesse. Il convient alors de procéder à des dénombrements ainsi que de contrôler les hommes. L'ordonnance de Villers-Cotterêts qui institue l'obligation de l'enregistrement des naissances, des morts et des baptêmes interdit aussi les coalitions ouvrières. En parallèle, l'enregistrement des baptêmes permettra à l'Église catholique de déceler les adeptes des autres religions (protestants, juifs…). Les recensements apparaissent donc comme un instrument privilégié du gouvernement suivant un des préceptes de Descartes « de faire partout des dénombrements si entiers et de revues si générales qu'il fut assuré de ne rien omettre ».

La pratique des dénombrements développe les réflexions théoriques sur les méthodes à mettre en œuvre. Le mouvement s'accentue au xviie et XVIIIe siècle, parallèlement à l'achèvement de la concentration du pouvoir entre les mains du monarque. La naissance de la statistique administrative en Allemagne, en France et en Angleterre illustre la diversité des approches[3]. Les stratégies de recherche d'informations économiques et sociales françaises opposent l'arithmétique politique anglaise à la statistique descriptive allemande.

Les statisticiens allemands raisonnent du point de vue de la puissance et de l'activité de l'État. La statistique consiste à recenser tout ce qu'il y a d'intéressant pour l'État, elle est descriptive et non quantitative. L'objet principal de l'effort statistique est de classer, d'organiser des observations hétéroclites. Le résultat est la construction de nomenclatures, un des aspects de la statistique moderne aussi essentiels que la dimension quantifiée.

Le contexte anglais est tout différent, il existe une société civile distincte de l'État qui est une partie de la société, et non sa totalité comme en Allemagne. Dans les années 1660, un ensemble de techniques d'enregistrement et de calcul apparaissent sous le concept de « L'arithmétique politique » qui en

3. Alain Desrosières, *La politique des grands nombres*, Paris : La Découverte, 1993 (Textes à l'appui).

utilisant les inscriptions dans les registres (des baptêmes, des mariages…) fournit des données démographiques et économiques pour l'ensemble du pays. Ce souci quantitatif s'applique tout d'abord aux domaines de l'économie et de la démographie. La conception libérale anglaise de l'État interdisant les grandes enquêtes, les arithméticiens doivent recourir à des méthodes indirectes de calcul.

Les deux conceptions de la statistique se confrontent puis un processus d'homogénéisation et de codification, l'unification des systèmes de référence, se réalise. L'opposition entre les deux approches se traduira et par la construction de nomenclatures – pour décrire – et par le recours au calcul – pour mettre en relation – au sein d'une production statistique administrative.

L'effort statistique se développe au cours du XIXe siècle avec une logique scientifique et dans une démarche intimement lié au politique : l'expansion des statistiques administratives est le signe d'une politique interventionniste, tandis que les périodes de régression sont les signaux d'une ère de libéralisme. Les crises et les bouleversements économiques auront pour effet une extension de l'intervention centralisée, à la fois sous la pression du mouvement ouvrier – contre l'anarchie du marché – et du sentiment des pouvoirs publics que l'absence de régulation automatique par le marché pouvait conduire à sa disparition c'est-à-dire aux révolutions. Les statistiques guident l'action des pouvoirs publics d'où les critiques récurrentes des libéraux non pas contre les chiffres, mais contre les statistiques[4].

Le besoin d'informations économiques s'explique du fait que, dans une économie décentralisée, les informations économiques ne sont pas immédiatement disponibles du fait de l'absence de liens institutionnels entre les producteurs et les consommateurs. L'explosion des statistiques au XIXe siècle, en Europe occidentale s'explique par l'expansion du capitalisme concurrentiel triomphant ; tout en se heurtant à l'opposition des libéraux pour qui les prix suffisent pour orienter les actions des agents. Dans le cours des transformations de l'économie, l'État prend de plus en plus de responsabilités dans l'économie, voire gère directement une partie des activités économiques, ce qui induit une forte croissance des besoins d'informations économiques et sociales et donc statistiques de la part des pouvoirs publics et des acteurs qui utilisent les mêmes indicateurs.

Le XIXe siècle s'est donné l'illusion de pouvoir tout connaître avec du temps et de l'énergie que ce soit la nature, l'homme ou la société. La connaissance statistique a participé de cette tentative et de cette illusion. Des auteurs

---

4. D'Adam Smith à Paul Fabra en passant par Jean-Baptiste Say et Friederich von Hayek, cf. « L'économie aveuglée » in *La cité des chiffres ou l'illusion statistique*, Paris : Autrement, 1992.

se préoccupent des relations entre la statistique et les sciences humaines, parmi ceux-ci le Belge Adolphe Quételet et le Français Augustin Cournot. Le premier se propose de connaître les phénomènes sociaux par leur unique aspect chiffré, la statistique appliquée aux actes humains devait constituer une science qu'il appelle physique sociale. Beaucoup de données statistiques publiées à l'époque nous paraissent actuellement sans intérêt. L'effort de mesure, de la recherche d'une évaluation de l'exactitude de la mesure conduit à développer la théorie des probabilités. Les méthodes de calcul s'affinent, la statistique s'affirme de plus en plus scientifique, le développement de la statistique continue en se complexifiant.

Le xxe siècle se caractérise par l'organisation de systèmes statistiques sous l'impulsion, voire sous la direction, des pouvoirs publics. La conception de l'information statistique se modifie. Sous la pression des groupes sociaux, l'information économique tend à être considérée comme un bien à la disposition du public. Le principe est donc la gratuité d'accès aux informations, sous réserve du coût du support. L'augmentation des demandes particulières en la matière conduit des organismes publics comme l'Institut national des études économiques (INSEE) à vendre l'information. Il s'agit là d'une inflexion importante qui modifie les conditions de diffusion de l'information statistique.

Le système statistique

Un système statistique se définit par un ensemble de pratiques, de méthodes et d'institutions, de ce point de vue, il existe un système statistique français. Ce système s'appuie sur l'existence de grands fichiers et répertoires d'industries ou d'établissements, des nomenclatures d'activités de métiers, des règles juridiques contraignantes. La réalité du système statistique est relativement récente.

> « Dans les faits, le système statistique français ne date que de la Libération, avec la création de l'INSEE, des services statistiques des ministères, des grands fichiers et répertoires d'individus et d'établissements, des grandes nomenclatures d'activités et de métiers et la systématisation des sondages. »[5]

Ce système est encadré par des règles juridiques contraignantes, en particulier le secret statistique.

Les systèmes statistiques contemporains naissent de la rencontre dans les années 1930-1940 de la statistique administrative – (la production des chiffres – et de la statistique scientifique – la production des méthodes.

5. Michel Lévy, *Comprendre les statistiques*, Paris : Seuil, 1979 p. 14.

Les prémisses apparaissent dans les années 1930 et les grands traits du système actuel se mettent en place à la Libération. Le système statistique français est le produit des besoins d'une gestion d'ensemble de l'économie nationale, spécialement de la planification, spécifique de la période de reconstruction de l'économie française. Il se constitue au cours de la Seconde Guerre mondiale puis s'organise et se développe durant la période de croissance des « trente glorieuses » (1945-1975), enfin il se transforme sous l'impact des grandes mutations contemporaines liées à la construction européenne, la mondialisation, etc.

La production des statistiques

Des données n'existent qu'en rapport à des objectifs. C'est pourquoi la production de données répond donc à des demandes explicites ou implicites des groupes sociaux, des administrations…, Si elles sont formalisées par les statisticiens, elles subissent des transformations et des modifications tenant autant aux choix des statisticiens eux-mêmes que des contraintes techniques ou sociales[6]. Les choix sont réalisés au niveau des décideurs, ensuite les professionnels mettent en œuvre les décisions prises. Toutes les informations ne seront pas produites, certains groupes sociaux influents peuvent bloquer la recherche ou la publication d'informations statistiques. Les statistiques disponibles dépendent du système statistique tel qu'il est organisé. Le producteur national dominant est l'INSEE dont l'essentiel des données se situe dans le cadre théorique de la comptabilité nationale.

Avant tout traitement, il faut s'assurer de la fiabilité des informations disponibles, sinon un calcul aussi complexe soit il n'aurait aucun sens. Les statistiques disponibles résultent d'un travail pratique mais elles sont aussi le produit de l'histoire. En effet, en statistique, il faut toujours garder à l'esprit que le présent n'est que l'aboutissement provisoire de processus historique. Le recueil des informations permet d'obtenir des données en grand nombre qu'il faut organiser pour les rendre utilisables.

Les informations proviennent soit d'une procédure explicite de recherche par le biais d'enquêtes ou de sondages soit d'une collecte et d'une mise en forme d'informations préexistantes. Dans tous les cas, la valeur des observations dépend étroitement des conditions du rassemblement des données brutes.

6. Voir en ce domaine l'ouvrage de Michel Volle, *Le métier de statisticien*, Paris : Hachette, 1980.

La recherche d'informations

L'enquête est une des méthodes courantes de recherche d'information, elle suppose d'avoir défini l'unité statistique enquêtée et la population de référence. Les recensements et les sondages sont les deux formes d'enquêtes réalisées directement auprès des détenteurs de l'information. Si les recensements sont exhaustifs, les sondages sont partiels. Certaines enquêtes combinent les deux techniques en fonction des sous-populations repérées. En statistique, un recensement est une étude exhaustive de toutes les unités statistiques de la population étudiée tandis que le sondage consiste à enquêter auprès d'une partie seulement de la population. Il faut faire attention car avec le temps, le terme de recensement, sous-entendu le recensement de la population d'un pays, a pris un autre sens courant. Il s'agit d'une enquête effectuée sur un échantillon de grande taille réalisée périodiquement et permettant une extrapolation des résultats qui constituent un « instantané » de la population d'un pays. Le recensement de la population en France est désormais organisé selon ce principe. L'ensemble des habitants des communes de moins de 10 000 habitants est recensé une fois tous les cinq ans par roulement. Dans les communes de plus de 10 000 habitants, chaque année une enquête est réalisée auprès de 8 % des habitants. Au bout de cinq ans tous les habitants des communes de moins de 10 000 habitants ont été recensés et 40 % des habitants des communes de plus de 10 000 habitants. Chaque année, les résultats de recensement sont produits à partir des cinq enquêtes annuelles les plus récentes.

Les recensements constituent l'opération statistique fondamentale. Le recensement n'est pas seulement un dénombrement, c'est aussi la mesure de certains caractères des individus de la population considérée. Le coût élevé des recensements en limite l'usage et la fréquence. Les recensements périodiques de la population constituent toujours une opération irremplaçable mais lourde et coûteuse.

Les sondages sont des enquêtes portant sur une fraction de la population. Cette technique est fondée sur le principe selon lequel les informations obtenues par l'interrogation d'un échantillon peuvent, sous certaines conditions, être généralisées à l'ensemble de la population. Cela peut consister à estimer certaines caractéristiques inconnues ou à faire des tests pour déterminer si des hypothèses ou des affirmations à propos de caractéristiques inconnues sont acceptables. La méthode des sondages présume l'existence de régularités au sein de la population concernée par les questionnaires. Un sondage tend à valoriser les modalités les plus courantes et à minimiser les signaux faibles significatifs. L'échantillon doit être construit de manière à rendre la

généralisation vraisemblable avec une bonne probabilité. Pour ce faire, deux méthodes sont possibles : les méthodes empiriques dont celle des quotas et les méthodes aléatoires ou probabilistes.

Il existe une autre méthode de l'échantillonnage dite de convenance qui consiste à interroger les individus sur le lieu d'achat ou d'activité ou dans la rue mais cette méthode, au petit bonheur la chance, n'offre aucune garantie scientifique. De plus, le sondage auprès de volontaires est un type de sondage qui demande à des individus de répondre : les téléspectateurs d'une chaîne particulière, ou les utilisateurs des réseaux sociaux, les lecteurs d'un journal. Là encore, aucune garantie scientifique ne peut être envisagée.

Dans la méthode des quotas, s'appuyant sur la connaissance de la répartition de la population selon des critères pertinents pour l'étude, il est possible de construire un échantillon représentatif possédant la même structure que la population mère. L'enquête doit respecter cette structure. Cette méthode est souple et rapide, mais le risque d'erreur est mal connu.

Les méthodes probabilistes désignent une technique par laquelle chaque unité de la population cible a une probabilité donnée, connue ou calculable préalablement (avant le tirage) d'appartenir à l'échantillon. Il devient alors possible de mettre en œuvre les techniques du calcul de probabilités pour réaliser des inférences sur l'ensemble de la population. Dans le sondage aléatoire simple, chaque individu de la population a une chance non nulle de faire partie de l'échantillon. Dans le sondage aléatoire stratifié, la population est subdivisée en strates, groupes homogènes selon un critère lié à la variable à estimer, par exemple les régions lors d'une enquête nationale. Puis au sein de chaque strate un échantillon est construit par sondage aléatoire simple. Dans le sondage en grappes, la population est divisée en plusieurs sous-ensembles, les grappes de l'échantillon sont choisies par sondage aléatoire simple et tous les individus appartenant aux grappes sélectionnées sont interrogés. Le sondage aréolaire sélectionne des aires au lieu de grappes. Dans un sondage à plusieurs degrés la population est divisée en grappes. L'échantillon est construit en tirant par sondage aléatoire simple au sein des grappes.

Les méthodes probabilistes exigent de disposer d'un recensement exhaustif de la population sous forme de répertoire comprenant la liste de toutes les unités. La constitution et la maintenance d'un répertoire sont des opérations complexes puisque toutes les unités sans exception doivent être répertoriées. La principale source d'erreur est constituée par l'absence de repérage d'unités. Pour les individus, seul l'INSEE dispose de cette base, mais en raison de la loi sur le secret statistique elle ne peut la communiquer à quiconque.

Le questionnaire est l'outil essentiel de toute enquête quantitative. Les questions retenues doivent permettre d'obtenir les informations recherchées. Pour cela, les personnes interrogées doivent être capables de répondre aux questions (elles doivent posséder les renseignements et estimer pouvoir les donner) et les comprendre sans ambiguïté. Les questions équivoques ou suggérant la réponse, comme c'est bien souvent le cas dans les sondages d'opinion, ne fournissent aucune information pertinente. Les résultats issus de ce genre d'enquêtes conduisent à des conclusions erronées. L'administration du questionnaire peut se faire par l'intermédiaire de différentes méthodes. Le mode d'administration le plus courant est le face-à-face entre l'enquêteur et le répondant. Actuellement, les tablettes ou les ordinateurs tendent à remplacer les questionnaires papier qui restent le mode le plus fiable d'administration. Les enquêtes via les réseaux sociaux de toute nature, outre les risques de discrimination qu'elles comportent, sont source de beaucoup de désillusions sur leur représentativité et leur fiabilité. Une fois les questionnaires recueillis, il faut en vérifier la cohérence et repérer les erreurs – il existe toujours des erreurs. Elles proviennent des enquêtés, dont les réponses peuvent être inappropriées (réponse imaginée par convenance) voire fausses (ressenti, mensonge…), des dispositifs d'obtention des informations (la représentativité des enquêtes via Internet est sujette à des interrogations, le recensement de 1968 fut particulièrement jugé peu fiable), du traitement de celles-ci (erreurs de codage), etc.

La collecte des informations existantes

Au cours de leurs activités pour les besoins de celles-ci, les agents économiques produisent des informations numériques. Les comptabilités d'entreprises en sont un exemple. Les statistiques des administrations, des ministères représentent une seconde source abondante de données. Tous ces sous-produits des activités des agents sont irremplaçables. Cependant, les données obtenues répondent aux besoins des agents non à ceux des économistes. Les réalités mesurées dépendent de définitions légales ou réglementaires et non de concepts ou notions économiques. Elles ne sont pas toujours adaptées aux objectifs des analystes d'où la nécessité de traduire ces informations dans les cadres adéquats.

Cette introduction fournit un cadre général de la démarche statistique, la suite de l'ouvrage développe les méthodes et techniques de traitement des données.

Organisation de l'ouvrage

Le premier chapitre présente les notions indispensables de la statistique descriptive et les représentations graphiques tandis que le deuxième donne les techniques utilisées pour l'analyse statistique des distributions à une dimension avec un accent mis sur les tendances centrales et les caractéristiques de dispersion. Le troisième chapitre fournit les outils classiques de l'étude des distributions à deux dimensions et de la mise en lumière des liens entre deux variables. Le chapitre suivant sera l'occasion d'une étude d'une forme particulière des séries à deux dimensions, dont le temps : ce sont les chroniques très présentes dans le champ de l'économie. Nous verrons enfin, dans un cinquième chapitre les indices, un des outils les plus indispensables et les plus controversés dans le champ de l'économie.

Chapitre 1

Les outils

Ce chapitre regroupe sous l'intitulé « Les outils » les concepts de base, les rudiments du vocabulaire des statistiques descriptives – les premières notions du calcul statistique – et enfin les principales représentations graphiques.

Les concepts de base

Avant tout calcul statistique, il est nécessaire de disposer de données. Pour atteindre cet objectif, il s'agit de définir très précisément la population sur laquelle s'effectue l'enquête ainsi que les variables à analyser. Le type de ces variables conditionne les traitements statistiques qu'il sera possible de leur appliquer.

La population et les unités statistiques

Dans le vocabulaire statistique, une population est un ensemble dont chaque élément est un individu ou une unité statistique. Les termes de population et d'individus sont employés aussi bien lorsqu'il s'agit d'un ensemble d'êtres humains : « la population résidente en France », « les salariés d'une entreprise », etc., que d'un ensemble d'objets inanimés : « la production automobile pour une année », « le stock des machines à une date donnée », et même d'ensembles abstraits ou des événements : « ensemble des jours d'une année », « la série du revenu national depuis vingt ans ». Chaque observation porte sur une unité statistique.
La population soumise à l'analyse statistique doit être définie avec précision afin que l'ensemble considéré soit déterminé sans ambiguïté de sorte qu'un individu quelconque puisse y être affecté sans incertitude. Quand on veut traiter de la population française au 1er janvier 2013 : il faut indiquer si

les étrangers résidant en France sont inclus et comment sont comptabilisés les Français résidants à l'étranger, il faudra alors préciser la signification du terme « résider ». Les chiffres de la population « française » correspondent aux nombres de personnes résidant en France métropolitaine et dans les départements d'outre-mer sous ces hypothèses, la population en France au 1er janvier 2013 est de 65,8 millions de personnes. Les habitants des collectivités territoriales d'outre-mer (Saint-Barthélemy Saint-Martin et Saint-Pierre-et-Miquelon Nouvelle-Calédonie, Polynésie Française, îles Wallis et Futuna) ne sont pas comptabilisés. De même les Français résidents à l'étranger, possédant parfois une double nationalité, ne sont pas décomptés. Par contre, les personnes de nationalité étrangère présentes sur le territoire appartiennent à la population de la France.
Comment définir les personnes employées dans une entreprise au 1er octobre 2013 ? Faut-il inclure les travailleurs à domicile, les travailleurs à temps partiel, les travailleurs intérimaires (salariés d'une autre entreprise), les stagiaires, les apprentis, les travailleurs « au noir » ? Doit-on comprendre les travailleurs absents pour maladie, en congé annuel ou en détachement ? L'effectif présent diffère en général de l'effectif théorique celui des personnes juridiquement salariées de l'entreprise.
Les règles qui définissent l'ensemble à étudier doivent permettre de dire sans ambiguïté si une unité appartient ou non au domaine. Pour chaque population statistique, il existe des conventions décrivant avec précision les limites de l'ensemble analysé. Il est exceptionnel que la population statistique et la population intuitive soient identiques.

Les caractères et les modalités

Pour décrire une population, on classe les individus selon certains attributs que l'on appelle des caractères (sexe, genre) ou des variables (âge). Il est indispensable de ne retenir que les caractères les plus pertinents pour pouvoir décrire une population convenablement. Il convient de ne sélectionner qu'un nombre restreint de caractères pour obtenir une description synthétique. Le caractère est un critère de classement, il peut présenter plusieurs situations différentes, il prend plusieurs modalités. Les deux modalités du caractère « sexe » sont : masculin et féminin. Ce caractère qui peut prendre deux modalités est dit dichotomique, une illustration de ce type de caractère : on peut « être au chômage » ou « ne pas être au chômage ». Le nombre de modalités d'un caractère dépend de l'information disponible et du but de l'étude. Par exemple, l'état matrimonial peut comprendre diverses modalités : célibataire, marié, veuf, divorcé, union libre, pacsé ou deux modalités :

cohabitant, non-cohabitant. Chaque individu de la population présente une et une seulement des modalités du caractère. Les modalités d'un caractère constituent une nomenclature, elles sont incompatibles et exhaustives, elles permettent de réaliser une partition de la population. Une unité statistique peut présenter plusieurs caractères, néanmoins son affectation sera fonction du caractère étudié. Tous les individus appartenant à un même sous-ensemble de la population sont équivalents du point de vue du traitement statistique.

Les caractères qualitatifs nominaux

Les caractères qualitatifs autrement appelés variables nominales ou variables catégorielles ont des attributs dont les différentes modalités ne sont ni mesurables ni repérables. Elles sont constatées. Avec l'usage de l'informatique, on utilise parfois le terme de données qualitatives. Le caractère se subdivise en catégories ou en modalités de la variable auxquelles seront attachés un effectif et une fréquence. C'est le cas pour le sexe, l'état matrimonial, la qualification professionnelle. Ce sont des noms ou des étiquettes permettant d'identifier une caractéristique de chaque élément. Même s'il n'est pas toujours possible d'y établir un ordre, les modalités sont rangées selon des critères plus ou moins arbitraires (Numéro Insee 1 pour les hommes, 2 pour les femmes). La présentation des modalités de la variable ne présuppose aucun classement.

Les caractères qualitatifs ordonnés

Certaines variables appellent naturellement un ordre dans le rangement de leurs catégories comme le niveau de formation. Elles sont repérables selon un type d'échelle plus ou moins légitime. Un caractère ordinal est un caractère qualitatif dans lequel les modalités possibles peuvent être classées dans un ordre spécifique ou dans un ordre naturel quelconque. Les catégories pourront alors donner lieu à un codage par les rangs qui ouvrira une autre gamme de traitements possibles proches de ceux des variables quantitatives. Dans le cas d'une nomenclature de formation, le classement est fonction du nombre théorique d'années d'étude nécessaires pour acquérir le niveau de formation. C'est de ce point de vue, une variable quantitative repérable. En effet, dans quelle mesure est-il légitime d'affirmer que le niveau I est supérieur au niveau III (comparaison d'un doctorat et d'un BTS) ?
Dans le tableau 1, le caractère « comportement » est ordinal parce que la modalité « Excellent » est meilleure que la modalité « Très bon », etc. La présentation est fondée sur un ordre « naturel », mais celui-ci est limité par le fait que l'information donnée ne permet pas de savoir quel élément du comportement différencie « Excellent » de « Très bon ».

Tableau 1. Classement des élèves selon le comportement.

Comportement	Nombre d'élèves
Excellent	5
Très bon	12
Bon	10
Mauvais	2
Très mauvais	1

Les variables textuelles

Une variable textuelle met en jeu des mots, des expressions langagières, voire des phrases qu'on ne peut réduire à des codes arbitraires, même si ceux-ci sont ordonnés. Il y a éventuellement un travail de préparation du texte, surtout s'il s'agit d'une transcription de textes oraux. En particulier, on peut lemmatiser, c'est-à-dire restreindre aux lemmes – passer en minuscule, au masculin singulier, à l'infinitif…

Une variable textuelle d'énonciation – ou semi-textuelle – ne met en jeu que des expressions traitées par comptage alors qu'une variable textuelle « complète » utilise des phrases, des segments et le calcul porte sur les mots, les lemmes ou les expressions à la fois des fréquences et des environnements. Ainsi la profession d'un adulte est une variable textuelle d'énonciation alors que la réponse à la question « pourquoi y a-t-il du chômage en France ? » est une variable textuelle « complète ».

La plupart des caractères qualitatifs requièrent une convention de définition ; c'est l'objet de la construction des nomenclatures.

Les caractères qualitatifs usuels et les nomenclatures

Les nomenclatures constituent des outils de classement des caractères qualitatifs. Loin d'être naturelles, elles résultent d'une réflexion scientifique et théorique y compris sur les modalités. Les différentes modalités d'un caractère constituent une nomenclature. La présentation des modalités d'un caractère qualitatif suit un ordre fondé sur un ensemble de travaux de recherche combinés avec les représentations usuelles de la population concernée.

Les différentes occurrences de la variable sont nominales, nous utiliserons le terme de modalité. Les différentes modalités d'un caractère constituent une liste appelée une nomenclature. Ces modalités peuvent être très nombreuses, leur regroupement en quelques grandes rubriques, généralement pas beaucoup plus qu'une dizaine, s'impose pour obtenir une description pertinente et utilisable ultérieurement du caractère considéré. Le choix de

ces rubriques, devenues les composantes élémentaires de la distribution, implique la mobilisation d'analyses théorique, statistique et sociale. Cette démarche longue et complexe assure l'acceptation de la nomenclature par les différents acteurs impliqués. La banalisation de l'usage des catégories socioprofessionnelles constitue un exemple d'une telle réussite.

Il existe un grand nombre de nomenclatures depuis celle des produits, d'activités, de catégories sociales ou de formation. Les nomenclatures font l'objet de négociations internationales au sein des grandes institutions ONU, OMC, FMI, Banque mondiale, OCDE. Elles sont souvent déclinées au plan régional avec les nomenclatures de l'Union européenne produites et gérées par Eurostat (http://epp.eurostat.ec.europa.eu/) qui dispose également de nomenclatures propres. L'élargissement des mesures statistiques à l'Europe a nécessité la création d'un système de codage harmonisé. Le service des statistiques des Communautés européennes Eurostat est l'Office statistique de l'Union européenne, dont la mission est de fournir des informations statistiques de qualité sur l'Europe. C'est le cas au plan français avec le Conseil national de l'information statistique (CNIS, http://www.cnis.fr/) qui assure la concertation entre les producteurs et les utilisateurs de la statistique publique. Les organismes publics de statistiques ont défini, dans un but de clarté et d'homogénéité, les catégories des variables qu'ils utilisent régulièrement. Les nomenclatures adoptées sont ensuite traduites en textes juridiques. Elles sont d'usages obligatoires au sein des administrations et recommandés pour les autres agents. Elles s'imposent de fait et deviennent d'usage banal comme la nomenclature des catégories sociales.

La nomenclature des professions et catégories sociales (PCS)

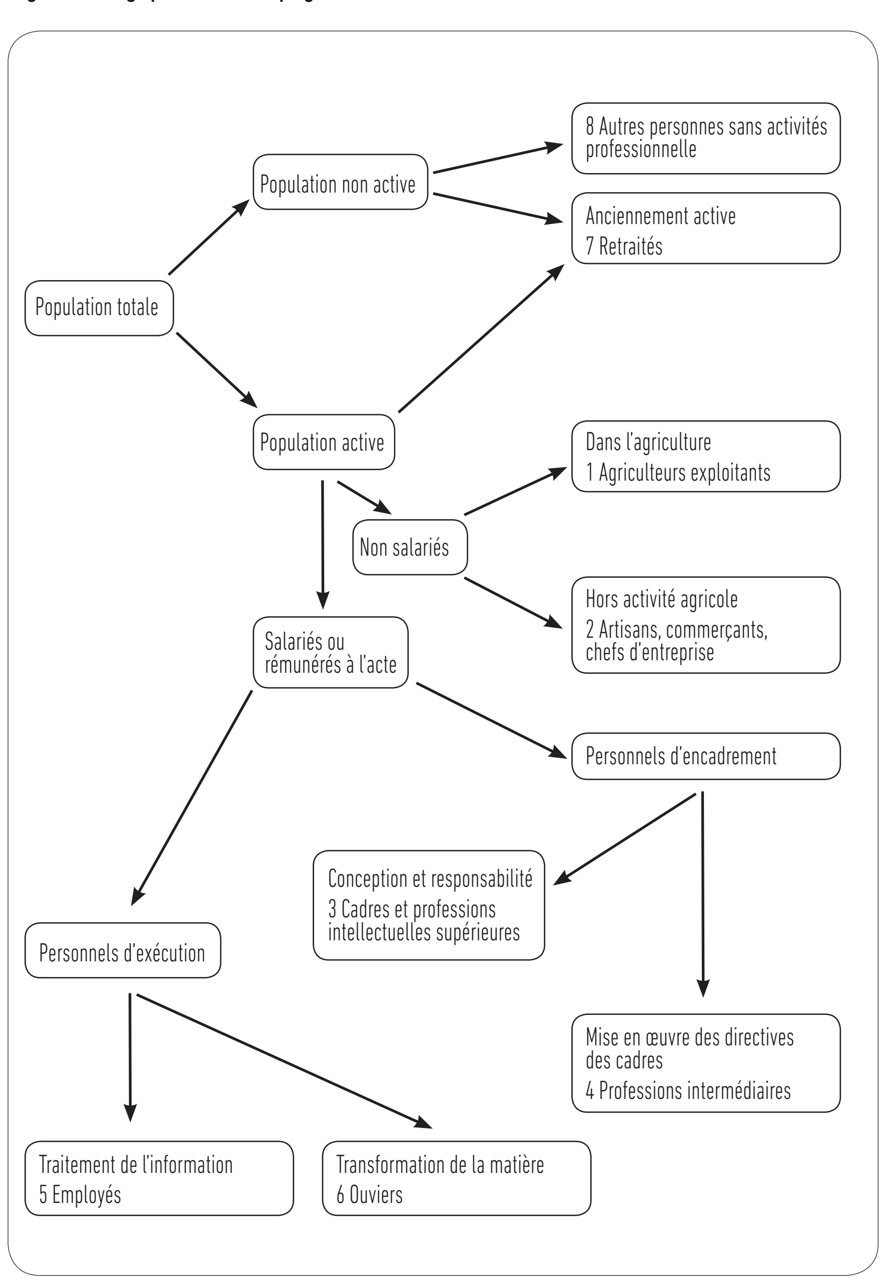
La nomenclature des PCS préserve les structures et les grands découpages de l'ancien code des catégories socioprofessionnelles. Le découpage résulte d'un ensemble d'études sur les comportements qui mettent en évidence une opposition entre les salariés des organismes producteurs de services essentiellement non marchands : le secteur public, et les salariés des entreprises qui ont des productions marchandes.

L'articulation des professions et des catégories socioprofessionnelles est assurée. Ces deux notions apparaissent comme deux niveaux d'une même nomenclature. Le premier chiffre du code indique le groupe socioprofessionnel (8 postes dont 6 pour les actifs occupés), les deux premiers chiffres indiquent la catégorie socioprofessionnelle dans les publications usuelles (24 postes), le niveau trois fournit les catégories détaillées (42 postes dont

32 pour les actifs occupés), le troisième niveau à quatre chiffres caractérise les professions (4 494 postes). Les publications courantes sont en 24 postes dont 19 pour les actifs.

Le schéma suivant décrit la logique de la nomenclature :

Figure 1. Logique du découpage.



Le tableau ci-après explicite les rubriques de la nomenclature.

Tableau 2. Les rubriques de la nomenclature.

Niveau agrégé[1] (8 postes dont 6 pour les actifs occupés)	Niveau de publication courante (24 postes dont 19 pour les actifs)
1 Agriculteurs exploitants	10 Agriculteurs exploitants
2 Artisans, commerçants et chefs d'entreprise	21 Artisans 22 Commerçants et assimilés 23 Chefs d'entreprise de 10 salariés et plus
3 Cadres et professions intellectuelles supérieures	31 Professions libérales 32 Cadres de la fonction publique, professions intellectuelles et artistiques 36 Cadres d'entreprises
4 Professions intermédiaires	41 Professions intermédiaires de l'enseignement, de la santé et de la fonction publique et assimilées 46 Professions intermédiaires administratives et commerciales des entreprises 47 Techniciens 48 Contremaîtres et agents de maîtrise
5 Employés	51 Employés de la fonction publique 54 Employés administratifs d'entreprises 55 Employés de commerce 56 Personnels des services directs aux particuliers
6 Ouvriers	61 Ouvriers qualifiés 66 Ouvriers non qualifiés 69 Ouvriers agricoles
7 Retraités	71 Anciens agriculteurs exploitants 72 Anciens artisans, commerçants, chefs d'entreprise 73 Anciens cadres et professions intermédiaires 76 Anciens employés et ouvriers
8 Autres personnes sans activité professionnelle	81 Chômeurs n'ayant jamais travaillé 82 Inactifs divers (autres que retraités)

1. Alain Goy, « La nouvelle nomenclature des professions et catégories socioprofessionnelles », *Courrier des statistiques*, n° 22 Avril 1982.

Ci-dessous, la Nomenclature d'activités françaises (NAF) constitue un exemple de nomenclature d'activités productives pour la France.

Tableau 3. Nomenclature d'activités française – NAF rév. 2, 2008.

NAF rév. 2, 2008 – Niveau 1 – Liste des sections	
Code	**Libellé**
A	Agriculture, sylviculture et pêche
B	Industries extractives
C	Industrie manufacturière
D	Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné
E	Production et distribution d'eau ; assainissement, gestion des déchets et dépollution
F	Construction
G	Commerce ; réparation d'automobiles et de motocycles
H	Transports et entreposage
I	Hébergement et restauration
J	Information et communication
K	Activités financières et d'assurance
L	Activités immobilières
M	Activités spécialisées, scientifiques et techniques
N	Activités de services administratifs et de soutien
O	Administration publique
P	Enseignement
Q	Santé humaine et action sociale
R	Arts, spectacles et activités récréatives
S	Autres activités de services
T	Activités des ménages en tant qu'employeurs ; activités indifférenciées des ménages en tant que producteurs de biens et services pour usage propre
U	Activités extraterritoriales

L'exemple ci-dessous illustre l'arborescence de la NAF et son articulation :
– M Activités spécialisées, scientifiques et techniques (Niveau 1 Sections)
– 71 Activités d'architecture et d'ingénierie ; activités de contrôle et analyses techniques (Niveau 2 Divisions)
– 71.1 Activités d'architecture et d'ingénierie (Niveau 3 groupes)
– 71.11 Activités d'architecture (Niveau 4 classes)
– 71.11Z Activités d'architecture (Niveau 5 sous-classes)

Cette nomenclature est la version française de la NACE (Nomenclatures statistiques des activités économiques dans la Communauté européenne) qui elle-même est une version de la nomenclature de l'ONU (CTCI – Classification type du commerce international).

La nomenclature de formation

Elle distingue les personnes selon leur niveau de formation le plus élevé en fonction des sorties du système de formation. La nomenclature française subdivise formation selon les rubriques suivantes : VI, Vbis, V, IV secondaire, IV supérieure, III, II et I.

- Les niveaux VI et Vbis correspondent aux sorties sans qualification.
- Niveau VI : sorties du premier cycle du second degré (6e, 5e, 4e) et des formations préprofessionnelles en un an (CEP, CPPN et CPA) ou des classes assimilées.
- Niveau Vbis : sorties de 3e, et des classes du second cycle court avant l'année terminale et des classes correspondantes de l'enseignement spécial.
- Niveau V : sorties de l'année terminale des seconds cycles courts professionnels (CAP, BEP) et abandons de la scolarité du second cycle long avant la classe terminale (seconde, première).
- Niveau IV : niveau des classes terminales du second cycle long.
  - Niveau IV secondaire : sorties des classes terminales du second cycle long.
  - Niveau IV supérieur : abandon des scolarisations post-baccalauréat avant d'atteindre le niveau III.
- Niveau III : sorties avec un diplôme de niveau bac + 2 (DUT, DEUG, BTS, formations sanitaires ou sociales).
- Niveaux II et I : sorties avec diplôme du second ou troisième cycle universitaire ou diplôme de grande école.

Exemple d'utilisation de cette nomenclature

Quel est le niveau de formation d'un étudiant qui, ayant suivi les cours de première année du DEUG de sociologie, n'a pas obtenu son passage en seconde année et quitte l'université ?
Le niveau de formation de cet étudiant sera IV, plus précisément IV sup. La nomenclature des formations est un caractère qualitatif ordonné.

Les variables quantitatives ou numériques

Les variables quantitatives ou variables statistiques ont des attributs dont les modalités sont exprimées sous forme numérique. Une variable quantitative

est soit mesurable soit repérable. À chaque unité statistique est associé un nombre : la valeur de la variable. Pour l'analyse statistique, il est habituel de distinguer les variables discrètes et les variables continues.

Variables numériques discrètes

Une variable dont les valeurs sont obtenues par dénombrement est une variable discrète. C'est par exemple le cas du nombre d'enfants. Une variable statistique est discrète ou discontinue lorsqu'elle ne peut prendre que certaines valeurs isolées – valeurs prises dans N plus rarement dans Z. C'est le cas du nombre de personnes qui composent un ménage. Un caractère discret peut prendre une infinité de valeurs dénombrables, il peut aussi n'en prendre que quelques-unes : le nombre d'enfants par familles qui est nécessairement un entier fini.
Certaines variables discrètes, comme le nombre de salariés d'une entreprise, pouvant prendre un très grand nombre de valeurs à l'intérieur d'un intervalle de grande amplitude, elles seront traitées comme des variables continues.

Variable statistique continue

Lorsque la variable peut prendre toutes les valeurs à l'intérieur d'un intervalle, la variable est dite quantitative continue par exemple la taille d'un individu, le revenu par habitant. Le nombre de modalités possibles est alors infini. La taille d'un individu est une variable continue, les revenus sont considérés comme continus, ce qui n'est pas tout à fait juste puisqu'ils ne peuvent prendre que des valeurs exprimées en centimes Les unités statistiques prenant sur ce type de variable un nombre très important de valeurs, il est nécessaire que les valeurs de la variable soient regroupées en classes.
Pour obtenir un nombre fini de modalités, les valeurs sont regroupées en classe. Les valeurs d'une variable continue sont mesurables ou repérables, avec un degré de précision déterminé qui n'est pas toujours connu pour les données économiques et sociales.
En pratique, la distinction entre variables discrètes et variables continues est conventionnelle. La précision d'une mesure est toujours limitée et les résultats seront toujours donnés sous forme d'un nombre fini d'observations. La production d'acier, par exemple, sera donnée en millions de tonnes ou en milliers de tonnes. Inversement, si une variable discrète peut prendre un grand nombre de valeurs, deux valeurs voisines apparaissent comme proches ; elle sera alors traitée comme une variable continue. La distinction repose sur le fait que les variables se présentent soit individualisées soit groupées en classe. En statistique, il est possible de perdre de l'information, en ce sens qu'il est possible de transformer une variable continue en variable discrète,

voire en caractère qualitatif. En revanche, bien que cela soit parfois pratiqué en attribuant une valeur numérique à une catégorie, l'opération inverse est plus problématique.

Exemple : Types de variable, variable ou caractère ?

- Quelle est la nature des caractères ci-dessous ?
- Nombre d'actions vendues chaque jour à la bourse
- Rémunérations des enseignants d'un lycée
- Indicateur du moral des ménages
- Écart de rémunération entre hommes et femmes
- Les pays de l'Union européenne
- Les niveaux de formation des salariés
- Les formes de contrat de travail
- Taux de croissance du PIB
- Prix à la consommation
- Solde commercial
- Indicateur du moral des ménages
- Nombre de personnes par ménages

Solution

- Nombre d'actions vendues chaque jour à la bourse variable discrète
- Rémunérations des enseignants d'un lycée variable quantitative continue
- Indicateur du moral des ménages variable qualitative ordonnée
- Écart de rémunération entre hommes et femmes variable continue
- Les pays de l'Union européenne caractère qualitatif
- Les niveaux de formation des salariés variable ordonnée
- Les formes de contrat de travail caractère qualitatif
- Taux de croissance du PIB variable quantitative
- Prix à la consommation variable quantitative
- Solde commercial variable quantitative
- Indicateur du moral des ménages caractère qualitatif ordonné codé
- Nombre de personnes par ménage variable statistique discrète.

Les classes

Les unités statistiques prenant sur ce type de variable un nombre très important de valeurs, il est nécessaire que les valeurs de la variable soient regroupées en classes avant tout traitement en un nombre fini de modalités. Le choix des classes répond en général aux exigences suivantes :

- elles ne doivent pas être trop nombreuses sinon il y aurait une difficulté de compréhension et de traitement ;
- elles ne doivent pas être trop peu nombreuses, car il y aurait perte d'information ;
- il ne doit pas y avoir de classe vide.

Le rangement des données, selon un ordre précis, est insuffisant dès que le nombre de données est grand. Pour étudier une variable continue, il faudra parfois regrouper les données sous une forme qui permette de ne pas perdre l'essentiel de l'information. Le regroupement ainsi effectué permet d'obtenir une distribution des fréquences ou des effectifs. Chaque classe aura un certain effectif ; certains auteurs utilisent le terme de fréquence absolue. Les calculs statistiques utiliseront les centres de classes comme représentatifs de l'ensemble de la classe. Les classes de valeurs possibles constituent les modalités du caractère étudié.

Le choix du nombre de classes et de leur amplitude est fonction de l'effectif de la population étudiée. Les classes peuvent avoir une amplitude variable ou constante. L'effectif de chaque classe ne doit pas être trop réduit pour éviter les fluctuations accidentelles. La variable « âge » est souvent subdivisée en classes d'amplitude de 5 ans, 0 moins de 5 ans, 5 ans moins de 10 ans, etc. 0, 5, 10 sont les extrémités des classes.

Pour rendre les calculs significatifs, tout en préservant la compréhension de la distribution, le nombre de classes doit être compris entre 5 et 15. Les classes doivent être agencées de telle sorte que toutes les informations soient incluses et que chaque observation se retrouve dans une et une seule classe. Les classes constituent ainsi une partition de l'ensemble considéré. Les amplitudes des classes ne doivent pas être trop différentes sauf pour les classes extrêmes.

La définition des classes

Les limites de classes doivent être sans équivoque. La présentation suivante est insatisfaisante.

Nombre de salariés par entreprises :

- 0 à 10
- 10 à 50
- …

En effet, il est impossible de classer sans équivoque les entreprises de 10 salariés qui peuvent appartenir à la première ou à la deuxième classe. Il en est de même pour les entreprises de 50 salariés.

D'autres possibilités sont utilisées comme :

- 0 à moins de 10 salariés
- 10 à moins de 50 salariés
- …

L'écriture la plus satisfaisante et la plus formelle est celle ci-dessous :

- [0, 10[
- [10, 50[
- …

Le nombre de classes à retenir dépend de la précision des mesures et de l'effectif de la population étudiée.

L'amplitude de classe

Le choix des amplitudes de classe est déterminé par le souci d'obtenir des effectifs comparables d'une classe à l'autre.
La valeur de l'amplitude d'une classe est calculée par la différence entre les valeurs de la borne supérieure et celle de la borne inférieure de la classe. Si la classe de rang i est $[b_i;b_{i+1}[$, l'amplitude a_i est définie par :

$$a_i = b_{i+1} - b_i$$

L'amplitude est donc pour la deuxième classe de [10, 50[est de 40. Il arrive que la borne inférieure de la première classe et la borne supérieure de la dernière ne soient pas données. Pour estimer ces bornes absentes, nous disposons de deux solutions. Tout d'abord réfléchir à ce que pourrait être la valeur de cette borne (ici pour la première classe 0 semble une solution satisfaisante). Sinon, il est courant d'affecter à la première classe l'amplitude de la seconde et à la dernière classe l'amplitude de l'avant-dernière, il faut cependant faire attention à ne pas avoir de données aberrantes.

Les centres de classe

Pour mener des calculs statistiques sur des séries classées, les classes sont réduites à une seule donnée : le centre de classe. L'opération consiste à transformer une distribution continue en distribution discrète en lui appliquant les outils applicables à une série continue. Cela revient à considérer que tous les individus d'une classe peuvent être décrits par ce centre de classe. Le centre de classe de la classe i est obtenu en prenant la moyenne des bornes de la classe i $[b_i;b_{i+1}[$ même si les bornes de la classe n'appartiennent pas nécessairement à la classe. Le centre de classe c_i se calcule simplement :

$$c_i = \frac{b_{i+1} + b_i}{2}$$

avec b_i la borne inférieure de la classe i et b_{i+1} la borne supérieure de celle-ci.
Par exemple pour la classe – [10,50[

$$c_i = \frac{b_{i+1} + b_i}{2} = \frac{50 + 10}{2} = 30$$

Les notions de base du calcul statistique

Ces notions constituent les fondements de la démarche statistique. C'est un ensemble de concepts, de principes et de méthodes indispensables donnant une signification aux résultats. Les chiffres significatifs, les signes conventionnels fournissent une première clef de lecture des données. Les formules statistiques recourent à un formalisme simple de valeurs indicées, de notations somme ou produit facilitant l'écriture et les démonstrations. La comparaison de deux ou plusieurs grandeurs utilise des outils comme les proportions, les fréquences, l'élasticité. La généralisation des calculs autorise la construction des tableaux des fréquences relatives et cumulées.

Les chiffres significatifs

Les résultats statistiques provenant de calculs le plus souvent réalisés à l'aide d'appareils de calcul numérique (calculettes, ordinateurs, tablettes…) s'expriment sous forme de nombre d'une grande précision. Il n'est pas rare de trouver des résultats avec trois ou quatre décimales. Une telle précision dégage un caractère de scientificité qui éteint toute critique, alors qu'il ne s'agit que d'une précision illusoire qui n'apporte aucune information. La précision des observations est telle que généralement les résultats sont donnés avec une seule décimale.
On appelle chiffres significatifs d'un nombre les chiffres exacts constituant ce nombre : 5,32 a trois chiffres significatifs. La précision du résultat ne doit pas être supérieure à la précision des observations. Le résultat final d'un calcul ne peut avoir plus de chiffres significatifs que le nombre entrant dans le calcul ayant le plus petit nombre de chiffres significatifs.

Exemple : $45{,}3 \cdot 65{,}326 = 2959{,}2678 \cong 2959{,}268 \cong 2959{,}27 \cong 2959{,}3$

L'écart entre le résultat « machine » 2959,2678 et le résultat retenu 2959,3 est de 0,0322 soit une erreur infime, très inférieure à la précision des données initiales. Pour autant cette procédure peut occasionner des imprécisions importantes si la variable est exprimée en millions ou en milliards.

Attention, pour les calculs intermédiaires tous les chiffres doivent être impérativement conservés.
Les pourcentages sont beaucoup utilisés dans les calculs statistiques. En général, compte tenu de la précision des données, le résultat final sera fourni avec une seule décimale.

Les signes conventionnels

Dans un tableau statistique, certaines informations sont absentes, remplacées par des signes conventionnels qu'il est utile de connaître.
« Le résultat n'existe pas faute d'enquête ou ne peut être obtenu
... Résultat non disponible (pas encore publié, pas encore parvenu)
/// Absence de résultat due à la nature des choses
— Résultat rigoureusement nul
c Résultat confidentiel par application des règles sur le secret statistique
ε Résultat inférieur à la moitié de l'unité choisie
e Estimation, évaluation
r Nombre rectifié
p Nombre provisoire
• Rupture de série

Les notations indicées

À chaque modalité, il sera possible d'associer un certain nombre d'individus, ce nombre sera appelé l'effectif de la modalité. Celui de la modalité i sera noté n_i . Soit k le nombre de modalités prises par un caractère ; nous noterons I l'ensemble des valeurs 1, 2,… k. L'ensemble constitué par les modalités et les effectifs associés à chacune des modalités forme une série statistique ou, plus usuellement, une distribution statistique du caractère pour la population considérée. Nous écrirons : $\{MO_i;n_i\}$ avec $i = 1,2,\cdots k$ où MO_i est la modalité i.

La notation somme (ou l'opérateur somme)

Soient les effectifs $n_1, n_2,\ldots, n_k$ de la distribution du caractère, nous noterons n la somme des effectifs.

$n = n_1 + n_2 + \cdots + n_k$

Cette écriture est peu maniable, nous remplacerons la somme précédente par la notation suivante :

$\sum_{i=1}^{k} n_i = n$ avec $i = 1,2,\cdots,k$. $(i \in N, k \in N)$,

ou si la sommation est sans ambiguïté : $\sum n_i = n$.

Le symbole Σ se lit *sigma* ou somme et il signifie la somme des effectifs des k modalités de la distribution. C'est un opérateur linéaire.

Quelques propriétés de cet opérateur :

$$\sum_{i=1}^{k}(x_i + y_i) = \sum_{i=1}^{k} x_i + \sum_{i=1}^{k} y_i$$

$$\sum_{i=1}^{k} a x_i = a\sum_{i=1}^{k} x_i$$

si a est une constante alors $\sum_{i=1}^{k} a = ka$

$$\sum_{i=1}^{k}(x_i + b) = \sum_{i=1}^{k} x_i + kb$$

$$\sum_{i=1}^{k}\ \sum_{j=1}^{l} x_{ij} = \sum_{j=1}^{l}\ \sum_{i=1}^{k} x_{ij} = \sum_{i=1}^{k}\left[\sum_{j=1}^{l} x_{ij}\right] = \sum_{j=1}^{l}\left[\sum_{i=1}^{k} x_{ij}\right]$$

$$\frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} y_i} = \sum_{i=1}^{k} \frac{x_i}{\sum_{i=1}^{k} y_i}$$

Attention car

$$\sum_{i=1}^{k} x_i^2 \neq \left(\sum_{i=1}^{k} x_i\right)^2$$

$4^2 + 3^2 = 25$ alors que $(4 + 3)^2 = 49$

$$\sum_{i=1}^{k} \sqrt{x_i} \neq \sqrt{\sum_{i=1}^{k} x_i}$$

$\sqrt{25} + \sqrt{49} = 5 + 7 = 12$ alors que $\sqrt{25 + 49} = \sqrt{74} = 8{,}6$

$$\sum_{i=1}^{k}\left(\frac{x_i}{y_i}\right) \neq \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} y_i}$$

$\frac{6}{5} + \frac{12}{7} = \frac{102}{35} = 2{,}9$ alors que $\frac{6+12}{5+7} = 1{,}5$

La notation produit (opérateur produit)

De façon analogue à la notation somme, nous écrivons le produit de n nombres sous une forme abrégée.

$$n_1 \cdot n_2 \cdot \ldots \cdot n_p = \prod_{i=1}^{p} n_i$$

$$\prod_{i=1}^{p} ax_i = a^p \prod_{i=1}^{p} x_i$$

$$\prod_{i=1}^{p} a = a^p$$

$$\prod_{i=1}^{p} x_i y_i = \prod_{i=1}^{p} x_i \prod_{i=1}^{p} y_i$$

Les pourcentages et les fréquences

Le calcul d'une proportion ou d'une fréquence est l'acte statistique le plus élémentaire. Cette simple opération donne déjà une information plus accessible que la distribution statistique, elle permet de comparer des distributions dont les ordres de grandeur ne sont pas comparables. Les pourcentages et les fréquences recouvrent des calculs formellement semblables.

Les proportions

Une répartition quantitative est le plus souvent exprimée sous forme de proportions. Une proportion indique l'importance relative d'une modalité dans l'ensemble des modalités. Une telle présentation permet de comparer des distributions statistiques dont les effectifs sont inégaux. Elle se calcule en faisant le rapport entre le nombre d'unités ayant le caractère et le nombre total d'unités.
Une forme très parlante de la présentation d'une proportion est de la formuler comme une fraction du numérateur 1/2, 1/3, 1/10. L'inconvénient d'une telle présentation est qu'il est malaisé d'effectuer des additions, il faut réduire les fractions à un dénominateur commun.
Pour simplifier les opérations, mais aussi pour permettre des comparaisons plus immédiates, il est courant de présenter les proportions avec un dénominateur commun 10 ou plus habituellement 100 car une proportion est le plus souvent donnée en pourcentage. Une proportion sera comprise entre 0 et 1, un pourcentage sera compris entre 0 et 100 %. Un pourcentage est une façon d'abréger les notations : $5\ \% = \frac{5}{100} = 0{,}05,\ 100\ \% = \frac{100}{100} = 1$.
Par exemple, en 2014, 18,5 % de la population française avait de 0 à 14 ans contre 22,2 % en 1981 et 19,3 % en 1991. Ce qui traduit un vieillissement relatif de la population de la France. Le résultat est arrondi en 2014, il y a 12 193 722 personnes de moins de 15 ans pour une population totale de 65 820 916 personnes. La proportion exacte est de 18,52560362423397 % en retenant un pourcentage de 18,5 la population des moins de 15 ans est de 12 176 869 moins de 15 ans soit un écart absolu de -16 852 personnes ou un écart relatif de 0,14 % par référence aux données de base. Cet écart est sans doute inférieur à l'erreur attachée aux résultats du recensement.

Le calcul d'un pourcentage consiste à appliquer le principe des proportions donc à poser l'équation suivante :

$\frac{x}{100} = \frac{a}{b}$, a est b étant connus, il découle :

$x = \frac{100 \cdot a}{b}$.

Les proportions sont parfois appelées des « taux » comme le soulignent les exemples qui suivent : Le taux d'activité est le rapport du nombre des actifs sur le total de la population concernée. Le taux de scolarité est la proportion de jeunes d'une population donnée suivant des études.
Une autre proportion est le plus souvent donnée en utilisant le vocable de taux. Le taux de chômage est le rapport entre la population active disponible à la recherche d'un emploi salarié et la population active. Le taux de chômage est le pourcentage de chômeurs dans la population active (actifs occupés + chômeurs). Le taux de chômage diffère de la part du chômage qui, elle, mesure la proportion de chômeurs dans la population totale. Cet exemple illustre une situation fréquente dans laquelle le numérateur fournit le sens du calcul alors que le choix du dénominateur est moins précis.
Les proportions expriment une situation au sein d'un ensemble fixe, l'appréciation des évolutions au cours du temps utilise la notion de taux.

La comparaison de l'évolution d'une valeur dans le temps ou dans l'espace : les taux

En économie, un taux mesure la modification relative d'une grandeur entre deux périodes. Il compare deux situations dans le temps. Soit Y une variable prenant les valeurs Y_0 et Y_1 aux temps t_0 et t_1. Le taux de croissance sera défini par : $r = \frac{Y_1 - Y_0}{Y_0}$ ou de façon plus générale $r = \frac{\Delta Y}{Y}$.

Pour exprimer la modification relative d'une grandeur, il est plus simple de l'exprimer à l'aide d'un multiplicateur ou d'un indice.
Nous pouvons écrire plus simplement :

$$r = \frac{Y_1 - Y_0}{Y_0} = \frac{Y_1}{Y_0} - 1 \; ou \; \frac{Y_1}{Y_0} = 1 + r \; ou \; Y_1 = Y_0(1 + r)$$

avec r le taux de croissance et 1 + r le multiplicateur.
Dans le cas de taux de croissance successifs, le calcul en sera facilité. Soit une croissance de r_1 suivie d'une de r_2, le multiplicateur de croissance est :

$$(1 + r) = (1 + r_1)(1 + r_2) = 1 + r_1 + r_2 + r_1 \cdot r_2$$

Donc le taux de croissance sur la période r est égal à :

$$r = (1 + r_1)(1 + r_2) - 1 = r_1 + r_2 + r_1 \cdot r_2 .$$

Du point de vue théorique, le taux de croissance global n'est pas égal à $r_1 + r_2$. Cependant, si le produit r_1 x r_2 est très petit devant $r_1 + r_2$, il est acceptable dans la pratique de calculer que le taux de croissance global est la somme des deux taux lorsque les deux taux initiaux sont faibles.
Exemple : soit une croissance de 15 % suivie d'une croissance de 10 %, le multiplicateur est 1,15 x1,10 = 1,265 soit un taux de croissance de 26,5 % et non de 25 %. (10+15). Par contre, pour un taux de 0,5 % suivi d'un taux de 0,4 %, le multiplicateur est : 1,005 x 1,004 = 1,00902 soit un taux de croissance de 0,902 % alors que la somme donne un taux de croissance de 0,9 %. La différence est dérisoire et le plus souvent non significative.
L'application à un ensemble de grandeurs économiques, des salaires par exemple, d'un taux de croissance identique, conserve les proportions, mais accroît les écarts absolus.

Tableau 4. Effets d'une croissance de 5 % sur les écarts de salaires mensuels.

Salaires		Écarts relatifs	Écarts absolus
1 260	3 240	2,57	1 980
Nouveaux salaires			
1 323	3 402	2,57	2 079

Une augmentation en valeur absolue conserve les écarts absolus, mais réduit les écarts relatifs.

Tableau 5. Effets d'une croissance de 100 € sur les écarts de salaires mensuels.

Salaires		Écarts relatifs	Écarts absolus
1 260	3 240	2,57	1 980
Nouveaux salaires			
1 360	3 310	2,46	1 980

Une croissance forte dans un pays pauvre peut se traduire par une augmentation de l'écart absolu avec un pays riche dont la croissance est incomparablement plus faible.

Le cas de l'Allemagne et de l'Argentine est intéressant.

Tableau 6. Effets de la croissance sur la richesse.

	PIB moyen 2009-2013	Taux de croissance	Multiplicateurs	Nouveaux PIB
	(milliards de dollars)	en % (hypothèse)		(milliards de dollars)
Allemagne	3 634 823	2	1,02	3 707 519
Argentine	611 755	10	1,1	672 931
Écart entre les PIB	3 023 067			3 034 588
PIB Argentin/PIB Allemand	16,8			18,2

Source : Banque mondiale

Malgré un taux de croissance cinq fois plus important, l'écart absolu et l'écart relatif de richesse s'accroissent entre l'Allemagne et l'Argentine. Un décryptage rapide des écarts croissants pourrait conclure à une détérioration de la situation en Argentine alors que la richesse disponible en Argentine a crû.

La comparaison de deux taux

Il est courant, en économie, de comparer l'évolution relative de deux taux : c'est le principe de l'élasticité. L'élasticité-prix mesure la variation relative de la demande d'un produit en réaction à une variation relative du prix de ce produit, elle s'exprime par le rapport de la variation relative de la quantité et de la variation relative des prix. Les deux mouvements sont, en général, de sens opposés ; l'élasticité est souvent négative.

$$e_p = \frac{\frac{\Delta q}{q}}{\frac{\Delta p}{p}} = \frac{\Delta q}{\Delta p}\frac{p}{q} = \frac{\Delta q}{q}\frac{p}{\Delta p}$$

– ep < 1 la demande est inélastique
– ep > 1 la demande est élastique
– ep = 1 aucune élasticité ou élasticité unitaire

Il existe d'autres formes d'élasticité, l'élasticité de l'offre $e_p = \frac{\frac{\Delta o}{o}}{\frac{\Delta p}{p}}$ (avec o la production, p les prix),

l'élasticité-revenu $e_p = \frac{\frac{\Delta revenu}{revenu}}{\frac{\Delta p}{p}}$ ainsi que des élasticités croisées $e_c = \frac{\frac{\Delta q_i}{q_i}}{\frac{\Delta p_j}{p_j}}$

qui mesurent les effets de l'évolution du prix d'un produit j sur la demande d'un produit i. La modification de la demande d'automobiles suite à l'augmentation du prix de l'essence en est un exemple classique. Une étude de l'INSEE de juin 2012 présente les sensibilités des dépenses de consommation des ménages en réaction aux modifications de prix et de revenus en utilisant les calculs d'élasticité.

Tableau 7. Élasticité-revenu et élasticité-prix par produits.

Élasticité-revenu Élasticité-prix	Non significativement différente de zéro	Faible (<0,5)	Proche de 1 mais <1	Forte >1
Faible (<0,5) ou non significativement différente de zéro	Textile-cuir	Produits alimentaires, Énergie, Services aux ménages	Transport	Matériels de transport, Services financiers
Importante (>0,5 et <1)		Autres produits industriels		Biens d'équipement Information-Communication
Forte (>1)			Commerce	Hébergement-Restauration

Marie-Emmanuelle Faure, Hélène Soual, Clovis Kerdrain, La consommation des ménages dans la crise, Note de conjoncture, Insee juin 2012

La généralisation des taux sous la forme de fréquences relatives facilite la compréhension des distributions.

Les fréquences relatives

Pour comparer les parts d'effectifs des différentes modalités, l'outil statistique le plus courant est constitué du calcul des fréquences relatives. L'avantage des fréquences par rapport aux effectifs est de pouvoir comparer des populations de tailles différentes ou mesurées par des unités différentes. La fréquence d'une valeur dans une série statistique mesure son importance relative, elle est le plus souvent exprimée en pourcentage. Elle se calcule comme l'importance d'une modalité par rapport à l'ensemble des modalités. Pour un caractère K ayant M_i modalités $1 \leq i \leq k (i \in N, k \in N)$, la fréquence de la modalité M_i sera notée f_i et se définit comme la proportion des individus de la population présentant la modalité M_i.

$$f_i = \frac{n_i}{n} = \frac{n_i}{\sum_{i=1}^{k} n_i}, \text{ avec } \sum_{i=1}^{k} f_i = 1$$

Une fréquence est toujours comprise entre 0 et 1 ou entre 0 et 100 % si elle est exprimée en pourcentage. Le total des fréquences est 1 ou 100 %.

Exemple

Tableau 8. Données économiques des opérateurs de jeux (2012).

	Mises	Fréquences relatives	Gains	Fréquences relatives
	en Mds €	f_i en %	en Mds €	f_i
FDJ	12,1	26,2	7,9	0,215
PMU	10,5	22,8	8,0	0,218
Casinos	15,4	33,4	13,1	0,357
Jeux en ligne (hors FDJ et PMU)	8,1	17,6	7,7	0,210
Total	46,1	100,0	36,7	1,000

Source : INSEE à partir des rapports d'activité des opérateurs de jeux, Esane, Arjel, DGFIP.

Il est possible de regrouper les fréquences relatives selon l'ordre de classement des séries ordonnées ou classées.

Les fréquences cumulées

Dans le cas de variables quantitatives et parfois dans celui des variables qualitatives ordonnées, les valeurs sont présentées par ordre croissant ou par ordre décroissant des fréquences. Il devient pertinent de calculer les fréquences cumulées en suivant la croissance ou la décroissance de la variable. Soit une variable statistique prenant différentes valeurs x_i ou c_i et les fréquences relatives f_i associées, la fréquence cumulée F_i est la somme des fréquences relatives des valeurs inférieures à x_i.

$F_1 = f_1$

$F_2 = f_1 + f_2$

$F_i = f_1 + f_2 + \cdots + fi$

F_i s'appelle la fréquence cumulée de la classe i, le recours au formalisme simplifie l'écriture.

$F_i = f_1 + f_2 + \cdots + fi$

$$F_i = \sum_{j=1}^{i} f_j$$

$$F_k = \sum_{j=1}^{k} f_j = 1$$

Les fréquences cumulées sont considérées comme les valeurs en des points connus d'une fonction de distribution F(x).

Tableau 9. Une distribution d'entreprises selon le nombre d'emplois.

Classes	Amplitudes de classe	Centres de classe	Effectifs	Fréquences (en %)	Fréquences cumulées (en %) croissante	Fréquences cumulées (en %) décroissante
	a_i	c_i	n_i	f_i	F_i	F_i
[0 ; 50[	50	25	177	40,1	40,1	100
[50 ; 100[	50	75	74	16,8	56,9	59,9
[100 ; 250[	150	175	84	19,0	75,9	43,1
[250 ; 500[	250	375	53	12,0	87,9	24,1
[500 ; 1000[	500	750	36	8,2	96,1	12,1
+ 1 000	500	1 250	17	3,9	100,0	3,9
Total			441	100,0		

La fréquence cumulée indique la proportion des individus statistiques, ici des entreprises, pour lesquelles la valeur de la variable statistique est inférieure à une valeur donnée. Par exemple, 75,9 % des entreprises emploient au plus 250 salariés. Ce qui revient à dire que 12,1 % des entreprises emploient au moins 500 salariés.

Les tableaux statistiques

Sauf cas exceptionnels, les données statistiques sont présentées sous forme de tableau. D'une part, cela permet d'appréhender l'information qui est synthétisée et d'autre part facilite ou rend possible les calculs. Ils constituent le moyen le plus sûr de pouvoir répondre aux questions posées de par leur systématisme.

Tableau 10. Tableau statistique pour une variable qualitative.

Modalités ou caractères	Effectifs	Fréquences	Pourcentages	Fréquences cumulées
	n_i	f_i	p_i	F_i
Catégorie 1	n_1	f_1	p_1	F_1
Catégorie i	n_i	$f_i = \frac{n_i}{n}$	$p_i = f_i \cdot 100$	$F_i = \sum_{k=1}^{i} f_k$
Catégorie m	n_m	f_m	p_m	F_m=1
	$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$	$\sum_{i=1}^{m} p_i = 100$	

Tableau 11. Tableau statistique pour une variable quantitative discrète.

Valeurs de la variable	Effectifs	Fréquences	Pourcentages	Fréquences cumulées
x_i	n_i	f_i	p_i	F_i
x_1	n_1	f_1	p_1	F_1
x_i	n_i	$f_i = \frac{n_i}{n}$	$p_i = f_i \cdot 100$	$F_i = \sum_{k=1}^{i} f_k$
x_m	n_m	f_m	p_m	F_m=1
	$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$	$\sum_{i=1}^{m} p_i = 100$	

Tableau 12. Tableau statistique pour une variable quantitative continue.

Classes	Centres des classes	Effectifs	Fréquences	Pourcentages	Fréquences cumulées
	c_i	n_i	f_i	p_i	F_i
$[b_i;b_{i+1}[$	c_1	n_1	f_1	p_1	F_1
$[b_1;b_2[$	c_i	n_i	$f_i = \frac{n_i}{n}$	$p_i = f_i \cdot 100$	$F_i = \sum_{k=1}^{i} f_k$
$[b_{m-1};b_m[$	c_m	n_m	f_m	p_m	$F_m=1$
		$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$	$\sum_{i=1}^{m} p_i = 100$	

Un tableau statistique à deux variables, ici avec des effectifs prend la forme d'un tableau dit de contingence.

Tableau 13. Tableau pour deux variables qualitatives.

Variable 2 / Variable 1	Modalité 1	Modalité j	Modalité p	Effectif marginal de la variable 1
Modalité 1	n_{11}	n_{1j}	n_{1p}	$n_{1.} = \sum_{k=1}^{p} n_{1k}$
Modalité i	n_{i1}	n_{ij}	n_{ip}	$n_{i.} = \sum_{k=1}^{p} n_{ik}$
Modalité m	n_{m1}	n_{mj}	n_{mp}	$n_{m.} = \sum_{k=1}^{p} n_{mk}$
Effectif marginal de la variable 2	$n_{.1} = \sum_{k=1}^{m} n_{k1}$	$n_{.j} = \sum_{k=1}^{m} n_{kj}$	$n_{.p} = \sum_{k=1}^{m} n_{kp}$	$n = \sum_{k=1}^{p} n_{.k} = \sum_{k=1}^{m} n_{k.}$

Exemple : nombre de personnes dans les ménages

Soit la distribution des ménages selon leur composition.

Tableau 14. Distribution des ménages en France en 2010.

Nombre de personnes du ménage	Pourcentage des ménages
1	34,0
2	33,1
3	14,5
4	12,2
5	4,6
6 et plus	1,7
Nombre total des ménages	27 106,5 ménages en milliers

http://www.ined.fr/fr/france/couples_menages_familles/menages_nombre_personnes/

Construire le tableau statistique en calculant la population comptée dans cette étude.

Solution

Un ménage est constitué des personnes occupant une même unité d'habitation. La construction du tableau statistique nécessite de calculer l'effectif de chaque catégorie de ménages. Il est obtenu en multipliant le nombre total des ménages par son importance relative. Les résultats ont été arrondis au millier d'unités près.

Par exemple, le calcul du nombre de ménages comprenant trois personnes est le suivant : Effectifs des ménages de trois personnes = 27106,5 x 0,145 = 3924,2.

Tableau 15. Effectifs des ménages suivant le nombre de personnes dans le ménage.

Valeurs de la variable	Fréquences	Pourcentages	Effectifs (milliers)
x_i	f_i	p_i	n_i
1 personne	0,34	34,0	9 216,20
2 personnes	0,331	33,1	8 964,20
3 personnes	0,145	14,5	3 924,20
4 personnes	0,122	12,2	3 308,40
5 personnes	0,046	4,6	1 234,80
6 et plus	0,017	1,7	458,70
Total	1,000	100,0	27 106,50

Pour calculer le nombre de personnes concernées par l'étude, nous devons faire une hypothèse sur la taille des ménages de la classe « 6 personnes et plus ». Dans le cas où nous considérerions que la taille moyenne de cette catégorie de ménages est de 6,5, nous obtiendrions une population de 61 306,4 milliers de personnes. Si nous retenons l'hypothèse de 7 personnes par ménage dans cette classe, nous avons une population de 61 535,7 milliers de personnes.

Tableau 16. Nombre de personnes : hypothèse 6,5.

Nombre de personnes du ménage	Nombre de ménages (en milliers)	Nombre de personnes (en milliers)
Ensemble	27 106,50	
1	9 216,20	9 216,2
2	8 964,20	17 928,4
3	3 924,20	11 772,6
4	3 308,40	13 233,6
5	1 234,80	6 174
6,5	458,70	2 981,55
Total	27 106,50	61 306,4

Tableau 17. Nombre de personnes : hypothèse 7.

Nombre de personnes du ménage	Nombre de ménages (en milliers)	Nombre de personnes (en milliers)
Ensemble	27 106,50	
1	9 216,20	9 216,2
2	8 964,20	17 928,4
3	3 924,20	11 772,6
4	3 308,40	13 233,6
5	1 234,80	6 174
7	458,70	3 210,9
Total	27 106,50	61 535,7

La différence dans l'estimation de la population totale selon les deux hypothèses est de : 61 535,7 - 61 306,4 = 229,3 milliers de personnes. C'est approximativement la taille de la population de Bordeaux.

Exemple : les appels téléphoniques

La facture détaillée des appels d'un abonné à un réseau téléphonique est connue sur la période deux mois.

Effectuez le regroupement en classes de ces données selon les trois variables continues. Vous veillerez à ce que les classes que vous avez choisies respectent les conditions qu'on attend d'un regroupement en classes.

- Plage horaire de l'appel
- Durée des appels
- Montant des appels.

Tableau 18. Détail des appels.

Date	Heure	Durée	Montant
jj.mm		mm:ss	Hors taxe en centimes d'euros
09.03	11 h 12	11 h 25	2,570
11.03	21 h 16	6 h 38	1,040
14.03	9 h 40	1 h 29	3,070
15.03	17 h 4	0 h 19	0,610
15.03	17 h 28	2 h 17	1,970
15.03	18 h 31	2 h 24	2,070
15.03	20 h 27	7 h 10	1,100
15.03	20 h 53	6 h 47	3,160
16.03	16 h 15	14 h 38	3,310
16.03	20 h 5	4 h 27	2,190
17.03	10 h 41	5 h 25	1,180
17.03	14 h 36	0 h 34	2,460
17.03	15 h 17	10 h 21	2,320
17.03	16 h 6	7 h 39	1,690
17.03	21 h 5	12 h 17	1,690
23.03	11 h 31	0 h 35	2,460
24.03	13 h 30	21 h 14	4,850
24.03	21 h 22	27:02	3,400
25.03	16 h 34	0 h 10	2,460
25.03	21 h 19	0 h 37	2,460
26.03	18 h 58	1 h 44	4,280
27.03	16 h 31	3 h 44	0,700
27.03	22 h 51	3 h 47	0,710
30.03	15 h 34	0 h 37	2,460
01.04	18 h 48	5 h 45	1,250
01.04	21 h 55	10 h 52	1,530

Date	Heure	Durée	Montant
jj.mm		**mm:ss**	**Hors taxe en centimes d'euros**
03.04	10 h 9	0 h 44	3,070
04.04	20 h 55	3 h 18	0,650
05.04	20 h 47	6 h 35	3,080
06.04	20 h 56	8 h	1,190
07.04	20 h 20	7 h 41	1,160
10.04	17 h 14	0 h 42	1,840
21.04	8 h 58	5 h 23	1,170
21.04	9 h 28	4 h 4	0,860
21.04	9 h 34	6 h 35	1,450
24.04	14 h 21	0 h 26	2,460
26.04	9 h 5	0 h 31	3,070
26.04	20 h 50	14 h 1	1,890
26.04	21 h 40	3 h 33	0,680
26.04	21 h 44	10 h 29	1,480
27.04	20 h 42	3 h 3	0,620
27.04	20 h 48	8 h 19	1,230
27.04	20 h 57	8 h 49	1,290
27.04	21 h 11	4 h 49	0,830
27.04	21 h 17	5 h 10	0,870
27.04	21 h 22	3 h 16	0,650
29.04	20 h 29	5 h 46	0,940
01.05	14 h 1	1 h 33	3,070
01.05	17 h 11	3 h 17	0,650
03.05	20 h 38	9 h 40	1,390
03.05	20 h 57	12 h 22	1,700
04.05	16 h 13	20 h 31	4,680
04.05	20 h 28	3 h 4	0,620
05.05	14 h 19	8 h 36	1,910
05.05	14 h 49	26:28	6,060
05.05	19 h 54	5 h 48	14,400

Corrigé

Pour effectuer le dépouillement sur la plage horaire de l'appel, nous constatons qu'aucun appel n'a lieu avant 8 h et aucun après 23 h. Nous choisissons des plages de deux heures démarrant à 8 h, la dernière plage sera une plage d'une heure.

Tableau 19. Plage horaire de l'appel.

Classes	Effectifs
	n_i
[8 ; 10[	5
[10 ; 12[	4
[12 ; 14[	1
[14 ; 16[	7
[16 ; 18[	9
[18 ; 20[	4
[20 ; 22[	25
[22 ; 23] [	1
Total	56

Avant d'aller plus loin dans le tableau, nous constatons un déséquilibre au niveau de la plage entre 20 h et 22 h. La perte d'information est importante et peut facilement être réduite en utilisant des plages d'une heure pour ce créneau horaire.

Tableau 20. Plage horaire de l'appel.

Classes	Centres des classes	Effectifs	Fréquences	Pourcentages	Fréquences cumulées
	c_i	n_i	f_i	$p_i\%$	F_i
[8 ; 10[	9 h	5	0,089	8,9	0,089
[10 ; 12[	11 h	4	0,071	7,1	0,161
[12 ; 14[	13 h	1	0,018	1,8	0,179
[14 ; 16[	15 h	7	0,125	12,5	0,304
[16 ; 18[	17 h	9	0,161	16,1	0,464
[18 ; 20[	19 h	4	0,071	7,1	0,536
[20 ; 21[	20 h 30	15	0,268	26,8	0,804
[21 ; 22]	21 h 30	10	0,179	17,9	0,982
[22 ; 23]	22 h 30	1	0,018	1,8	1,000
Total		56	1,000	100,0	

Tableau 21. Durée des appels.

Classes	Effectifs
	n_i
[0 ; 5[	26
[5 ; 10[	16
[10 ; 15[	10
[15 ; 20[	0
[20 ; 25[	2
[25 ; 30]	2
Total	56

Avant d'aller plus loin, deux constats : les premières classes sont disproportionnées par rapport aux autres et il y a une classe vide. Un autre découpage possible est explicité dans le tableau suivant.

Tableau 22. Durée des appels (regroupement plus judicieux).

Classes	Centres des classes	Effectifs	Fréquences	Pourcentages	Fréquences cumulées
	c_i	n_i	f_i	p_i%	F_i
[0 ; 2[	1 h	13	0,232	23,2	0,232
[2 ; 4[	3 h	10	0,179	17,9	0,411
[4 ; 6[	5 h	9	0,161	16,1	0,571
[6 ; 8[	7 h	7	0,125	12,5	0,696
[8 ; 10[	9 h	5	0,089	8,9	0,786
[10 ; 12[	11 h	4	0,071	7,1	0,857
[12 ; 14[	13 h	2	0,036	3,6	0,893
[14 ; 16[	15 h	2	0,036	3,6	0,929
[16 ; 30[	23 h	4	0,071	7,1	1,000
Total		56	1,000	100,0	

Tableau 23. Montants des appels.

Classes	**Centres des classes**	**Effectifs**	**Fréquences**	**Pourcentages**	**Fréquences cumulées**
	c_i	n_i	f_i	$p_i\%$	F_i
[0 ; 1[	0 h 30	13	0,232	23,2	0,232
[1 ; 2[	1 h 30	20	0,357	35,7	0,589
[2 ; 3[	2 h 30	10	0,179	17,9	0,768
[3 ; 4[	3 h 30	8	0,143	14,3	0,911
[4 ; 5[	4 h 30	3	0,054	5,4	0,964
[5 ; 10[	7 h 30	1	0,018	1,8	0,982
[0 ; 30[	22 h 30	1	0,018	1,8	1,000
Total		56	1	100	

Exercice : analyse statistique d'une table de mortalité

Tableau 24. Extrait de la table de mortalité de la génération féminine française de 1899.

Âge exact	**Survivants à l'âge exact**
0	100 000
1	84 883
2	82 247
3	80 843
4	79 995
5	79 186
6	78 763
7	78 411

Source : « La mortalité par génération en France depuis 1899 », Travaux et documents, Cahier INED n° 63, 1 973

Présentez le tableau statistique de la variable « âge du décès » sous sa forme habituelle. Donnez la signification concrète de chacune des colonnes du tableau statistique obtenu.

Corrigé

L'étude porte sur une population filles nées en 1899 et décédées avant l'âge de huit ans.

Comment obtient-on les effectifs du tableau ? Dans le cas de la première classe, il y a eu 100 000 naissances, un an plus tard seules 84 883 filles sont encore vivantes, le nombre de décès est donc de 100 000 - 84 883= 15 117. Le raisonnement répété pour tous les âges permet de construire le tableau statistique.

Tableau 25. Table de mortalité de la génération féminine française de 1899.

Classes	Centres des classes	Effectifs	Fréquences	Pourcentages	Fréquences cumulées
	c_i	n_i	f_i	p_i	F_i
[0 ; 1[	0,5	15 117	0,700	70,0	0,700
[1 ; 2[	1,5	2 636	0,122	12,2	0,822
[2 ; 3[	2,5	1 404	0,065	6,5	0,887
[3 ; 4[	3,5	848	0,039	3,9	0,927
[4 ; 5[	4,5	809	0,037	3,7	0,964
[5 ; 6[	5,5	423	0,020	2,0	0,984
[6 ; 7]	6,5	352	0,016	1,6	1,000
		21 589	1,000	100,0	

La première colonne reprend l'âge des décès.

La deuxième représente l'âge moyen du décès par classe annuelle.

La colonne des effectifs donne le nombre de filles décédées dans la tranche d'âge considérée. Au total, 21 589 filles sont décédées avant l'âge de 8 ans.

La colonne suivante donne la fréquence des filles décédées dans la tranche d'âge, ainsi que la cinquième qui exprime la même chose en pourcentage.

La dernière colonne donne la fréquence des filles décédées avant la borne supérieure de la tranche d'âge.

La fréquence F_i = 0,927, signifie que 92,7 % des filles mortes avant l'âge 8 ans sont mortes avant l'âge de 4 ans. Le tableau montre que la mortalité des fillettes est très forte au cours de la première année (70 % des décès).

Les représentations graphiques

Les graphiques permettent de donner une synthèse visuelle de la distribution d'une variable et de percevoir l'éventuelle relation entre les variables, cette section en présente quelques exemples. Les représentations peuvent être spécifiques à un type de variable ou de caractère. Sauf indication contraire, tous les graphiques sont réalisables en effectifs ou en fréquences, ils sont superposables à l'échelle près.

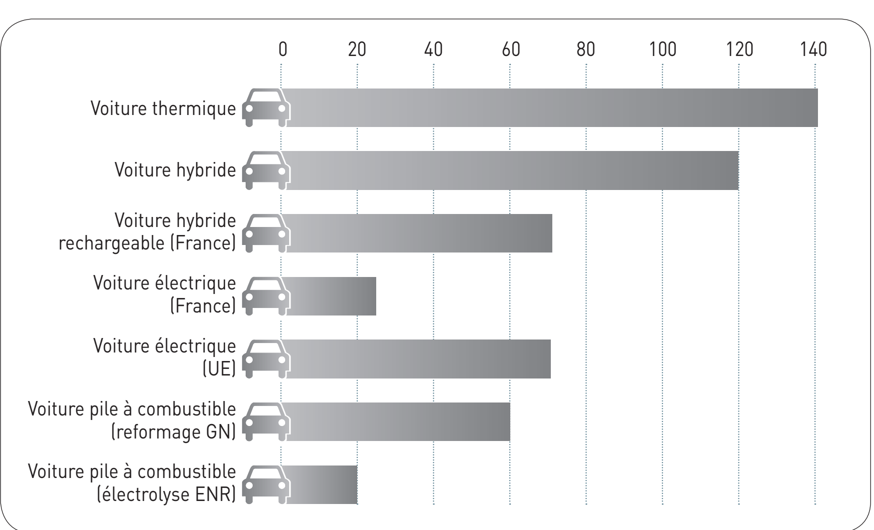
Le choix des représentations graphiques dépend pour une large part du type du caractère statistique : caractère qualitatif, variable statistique discrète ou variable statistique continue.

Les représentations des caractères qualitatifs

Les diagrammes figuratifs ou pictogrammes

D'une lecture aisée, les diagrammes figuratifs, les pictogrammes sont utilisés pour leur effet suggestif. L'importance relative du phénomène est figurée par des surfaces à partir de représentations imagées : un personnage pour une population humaine, des épis pour une production céréalière, des automobiles pour les émissions de gaz de serre.

Figure 2. Graphique en bâtons figuratifs : émission de CO_2 (g/km).

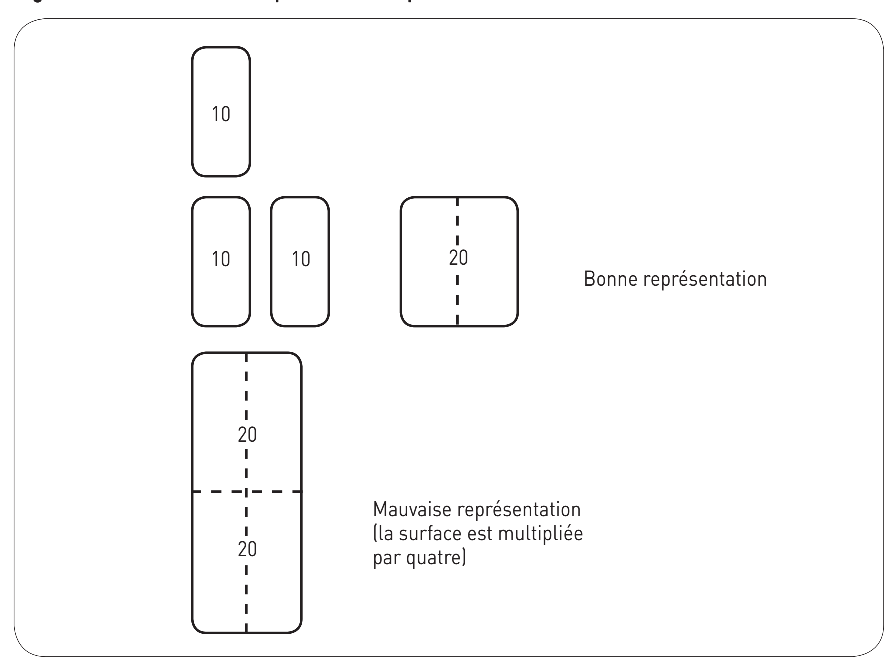


Source de ces données chiffrées : valeurs moyennes issues de diverses publications : études du Department of Energy américain et de l'IFP

Les illustrations utilisées pour figurer la distribution de caractère qualificatif sont souvent imprécises. Le lecteur ne sait pas toujours s'il faut comparer les longueurs ou les surfaces. Pour qu'un diagramme figuratif soit signifiant, il faut que les surfaces soient proportionnelles aux effectifs ou aux fréquences. Sinon, l'impression produite sera fausse. L'erreur la plus fréquente est de doubler les dimensions de la figure pour indiquer un doublement de la variable, alors que c'est uniquement la surface qui doit être multipliée par deux. La multiplication par deux des dimensions du diagramme indique une multiplication par quatre de la grandeur représentée.

L'exemple ci-dessous illustre cette situation.

Figure 3. Schéma d'une représentation par des surfaces.



Un exemple de cette représentation est l'évolution du pouvoir d'achat du dollar canadien.

Figure 4. Pouvoir d'achat du dollar canadien, de 1980 à 2000.



Cette représentation n'est intéressante que si les différences sont sensibles, doublement ou triplement. Une augmentation de 20 % d'une surface est peu lisible. De plus, la comparaison d'aires est, sauf longue habitude, difficile surtout si elles n'ont pas la même forme. Ces pictogrammes n'ont aucune ambition particulière ; ils n'apportent que peu d'information, sauf à mettre en exergue des classements.

Exemple : bilan des apports et des usages de l'eau : commentaire d'un pictogramme

Figure 5. Bilan des apports et des usages des eaux continentales (en milliards de m³/an).

Source : TEF 1998/1999, Paris : Insee

À partir de ce pictogramme, il sera possible de répondre à plusieurs questions : quelle est la quantité nette d'eau consommée par les centrales électriques ? Quel est le bilan des échanges d'eau avec les pays voisins ? Quelle est l'équation d'équilibre des usages humains de l'eau ? Quelle est l'équation d'équilibre des eaux continentales ?

Corrigé

Pour calculer la part des centrales électriques dans la quantité nette d'eau utilisée, nous supposons que la « consommation » nette, en fait l'évaporation, est proportionnelle à la quantité utilisée. Cette hypothèse est sans doute une sous-estimation dans le cas des centrales électriques. Pour un usage total de 41 milliards de m³ par an, 35 provenant des précipitations et 6 des nappes phréatiques, nous obtenons une consommation de :

$$6 \cdot \frac{26}{41} = 3{,}8 \textit{ milliards de } m^3$$

La France reçoit 11 milliards de m^3 par écoulement, les écoulements vers les pays voisins s'élèvent à 18 milliards de m^3 donc le solde des échanges extérieurs est de - 7 milliards de m^3.
Les activités humaines prélèvent 35 milliards de m^3 dans « les rivières, lacs et retenues » et 6 milliards de m^3 dans les nappes phréatiques, soit 41 milliards de m^3.
Les usages humains se traduisent par une évapotranspiration de 6 milliards de m^3, un écoulement de 4 milliards de m^3 vers la mer et de 31 milliards de m^3 dans les « rivières, lacs et retenues », soit également 41 milliards de m^3.

Tableau 26. Emplois et ressources des usages humains (milliards de m^3).

Emplois		Ressources	
Consommations nettes (évaporation)	6	Prélèvements dans les « rivières, lacs et retenues »	35
Écoulement vers la mer	4	Ponctions dans les nappes phréatiques	6
Rejets vers les « rivières, lacs et retenues »	31		
Total	41	Total	41

Les apports sont de 480 milliards de m^3 de précipitations (pluie, neige) et - 7 milliards de m^3 provenant des échanges avec les pays voisins. Les apports sont de 473 milliards de m^3.
Les usages comprennent 321 milliards de m^3 sous forme d'évapotranspiration, 150 milliards de m^3 sous forme de ruissellement et 2 milliards de m^3 d'écoulement des nappes phréatiques vers la mer. Nous obtenons bien 473 milliards de m^3.

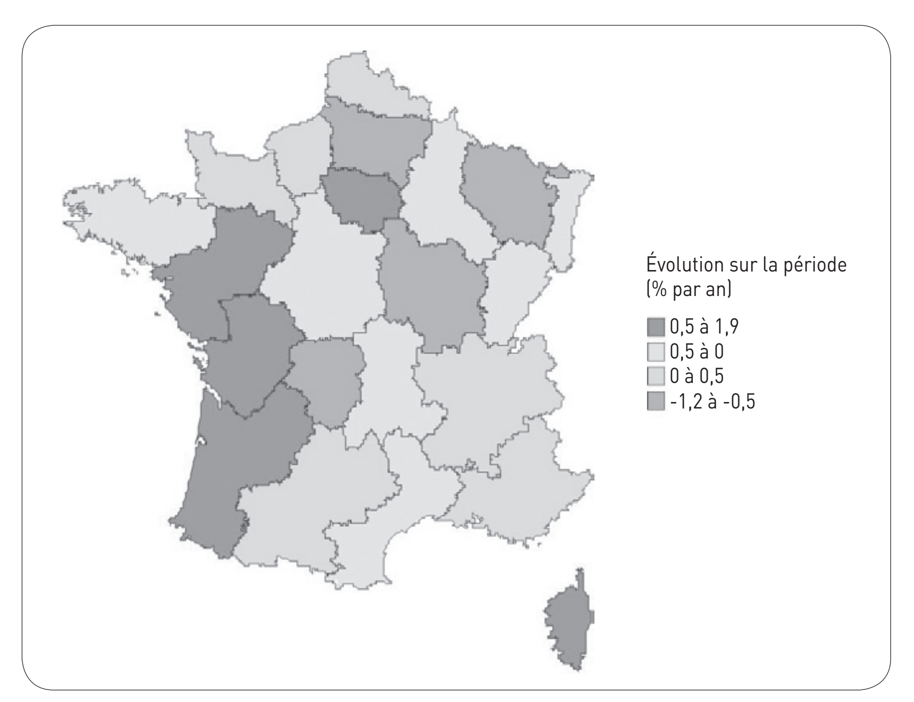
Tableau 27. Emplois et ressources des eaux continentales (milliards de m^3).

Emplois		Ressources	
Évapotranspiration	321	Précipitations	480
Ruissellement	150	Échanges avec les pays voisins	-7
Écoulement vers la mer	2		
Total	473	Total	473

Les cartogrammes

Eux aussi clairs et lisibles, les cartogrammes représentent les valeurs ou variations d'une grandeur sur un territoire géographique en assignant à chaque zone – département, région – ses caractéristiques. Pour cela, on utilise des fonds de cartes pour représenter les variables. Il existe deux grandes catégories de cartogrammes. Dans la première catégorie, les surfaces de chaque unité géographique sont par une gamme de hachures ou de couleurs propres à chaque classe du phénomène hachurées ou coloriées. L'impression retirée par le lecteur dépend à la fois de l'intensité des hachures ou des couleurs et de l'aire concernée.

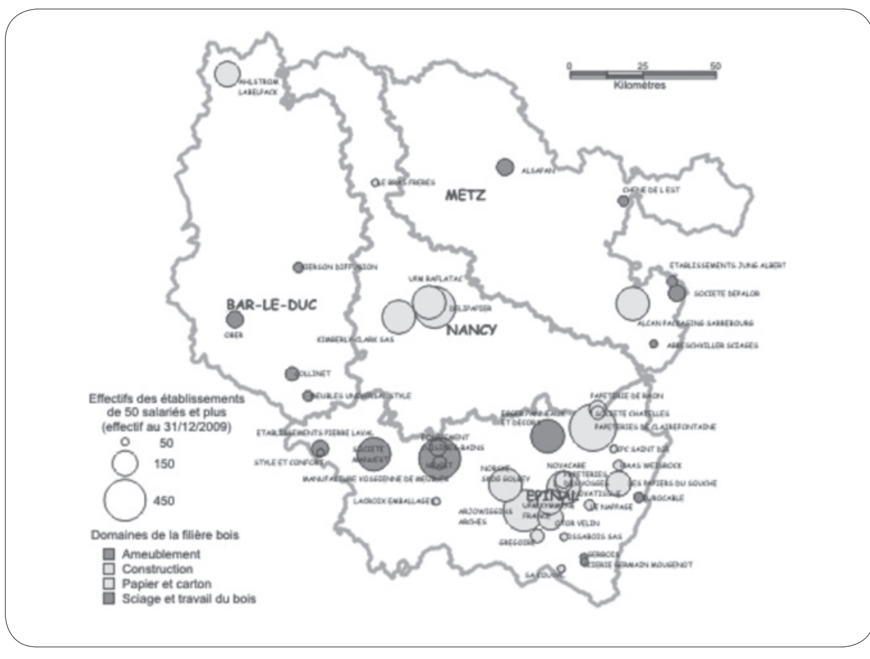
Figure 6. Évolution des PIB régionaux en volume entre 2008 et 2011.



Source : Insee, Comptes régionaux base 2005.

Dans la seconde catégorie, les phénomènes sont représentés par des surfaces proportionnelles centrées sur les unités géographiques et proportionnelles aux effectifs étudiés.

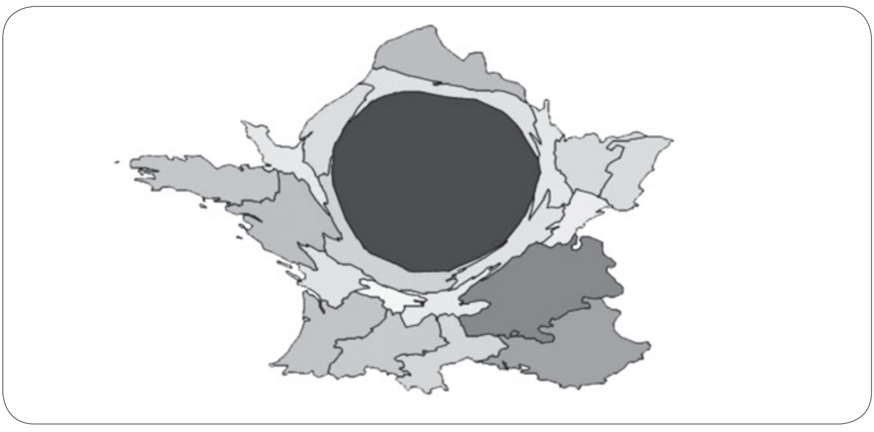
Figure 7. Les principales entreprises de la filière bois en Lorraine.




Emplois et entreprises du bois en Lorraine : une filière bien implantée
Pierre-Yves Berrard, Insee Lorraine, Noël Spitz, DRAAF Lorraine

L'utilisation d'outils de calcul très puissants peut fournir des représentations très évocatrices de l'importance d'un phénomène économique lié à une situation géographique.

Figure 8. La France en fonction des PIB régionaux.



© Journal du Net

Les diagrammes en colonnes

Le diagramme en tuyaux d'orgue, en barres ou en colonnes est constitué d'une suite de rectangles dont les hauteurs sont proportionnelles à l'effectif (ou à la fréquence) de la variable et dont les bases sont identiques. La représentation peut être horizontale ou verticale.

Exemple : les catégories socioprofessionnelles

Le tableau suivant fournit des informations sur l'importance des professions et catégories sociales dans la population française en 2003 et en 2012.

Tableau 28. Catégorie socioprofessionnelle (PCS).

	2003	2012
Agriculteurs exploitants	1,5	1,0
Artisans, commerçants, chefs d'entreprise	3,3	3,4
Cadres, professions intellectuelles supérieures	7,8	9,6
Professions intermédiaires	12,6	13,3
Employés	16,1	16,0
Ouvriers (y compris agricoles)	13,6	12,4
Inactifs ayant déjà travaillé	29,3	26,5
Autres sans activité professionnelle	15,8	17,7
Total	100,0	100,0

Champ : population des ménages de 15 ans ou plus, vivant en France métropolitaine. Résultats en moyenne annuelle. Source : Insee, enquêtes Emploi.

Définissez en quelques lignes ces différentes catégories.
Représentez l'évolution des catégories entre les deux dates.

Corrigé

Les différentes catégories sociales correspondent aux catégories de la nomenclature des PCS (Professions et catégories sociales) sont les suivantes :

– la catégorie des « Agriculteurs exploitants » regroupe les personnes dont l'activité principale est l'agriculture et qui ne sont pas salariées ;

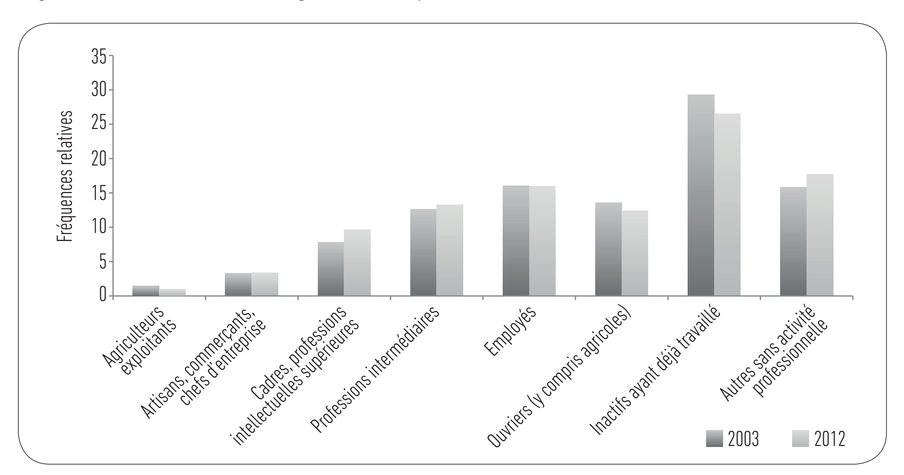
– les « Artisans, commerçants et chefs d'entreprise » regroupent les non-salariés n'ayant pas d'activités agricoles ;

– les « Cadres et professions intellectuelles supérieures » regroupent les salariés qui occupent des positions professionnelles supposant un niveau de formation au moins égal à la licence ;

– les « Professions intermédiaires » rassemblent des salariés qui sont en position intermédiaire ou ont des positions d'intermédiaires, la formation les « Employés » sont des agents d'exécution qui traitent et manipulent de l'information, des symboles ;
– les « Ouvriers » sont des agents d'exécution qui transforment la matière, la catégorie comprend les ouvriers agricoles ;
– les Inactifs ayant déjà travaillé sont les retraités ;
– les Autres sans activité professionnelle correspondent à des personnes n'appartenant pas à la population active.

Un diagramme en colonnes permet de visualiser l'importance et les évolutions des différentes catégories sociales.


Figure 9. Évolution des catégories socioprofessionnelles entre 2003 et 2012.



Le diagramme à accumulation interne ou à barres superposées permet de représenter plusieurs caractères au sein d'une même colonne ou d'un même graphique par exemple pour une période donnée.

Exemple : l'évolution des émissions de gaz à effet de serre (GES)

L'évolution des émissions de GES en France est donnée dans le tableau suivant par type.

Tableau 29. Les émissions de gaz à effet de serre en France.

En millions de tonnes d'équivalent CO_2	1990	1995	2000	2005	2008	2009
CO_2	394	393	409	420	391	373
CH_4	67	68	68	66	66	65
N_2O	93	91	78	68	66	62
HFC, PFC, SF_6	10	8	12	15	16	16
Ensemble	564	560	567	569	539	516

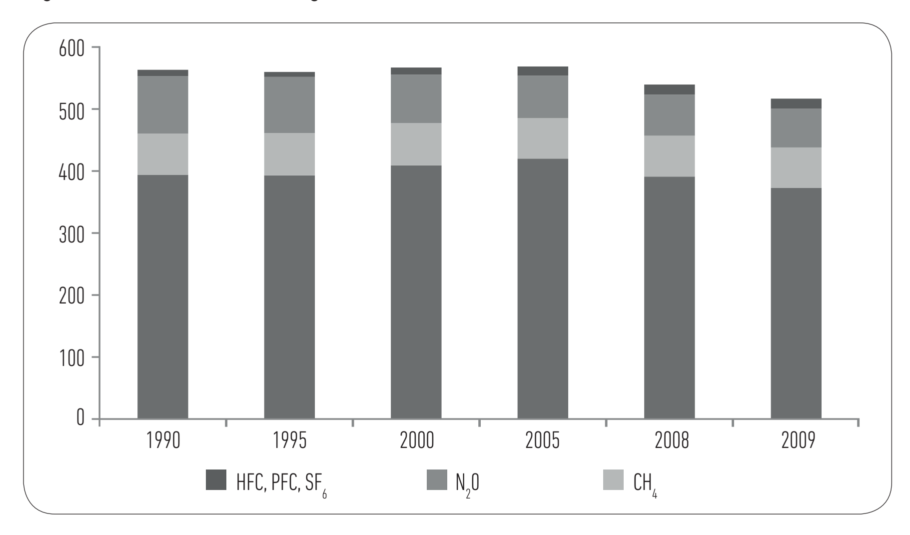
Source : Citepa (format CCNUCC), mai 2011.

Représentez graphiquement l'évolution de l'importance relative des divers GES.

Solution

Un graphique à accumulation interne permet de représenter à la fois l'importance relative des différents gaz et leurs évolutions au cours du temps. Les émissions mesurées sont celles de dioxyde de carbone (CO_2), de méthane (CH_4), de protoxyde d'azote (N_2O), d'hexafluorure de soufre (SF_6), d'hydrofluorocarbures (HFC) et de perfluorocarbures (PFC).

Figure 10. Les émissions de gaz à effet de serre en France.



Ce graphique illustre les transformations de la structure des émissions de GES. Il indique également la réduction globale des émissions de GES. Cette réduction paraît sensible en particulier pour le dioxyde de carbone (CO_2) et de protoxyde d'azote (N_2O).
Le diagramme en secteurs ou en « camembert » visualise la part relative des catégories de la variable sur une population. Le disque représente l'ensemble de la population, les différentes modalités seront représentées par des secteurs dont la surface est proportionnelle aux effectifs ou aux fréquences. Une telle représentation n'est significative que si le total des fréquences est de 100 %. Un demi-cercle peut jouer le même rôle.

Répartition des professions et catégories sociales

Une première représentation des professions et catégories sociales sous forme de diagramme en colonne a été utilisée ci-dessus. Les surfaces de la représentation circulaire rendent plus explicites les différences entre les différents groupes.

Exemple : PCS les proportions

Tableau 30. Catégorie socioprofessionnelle (PCS).

	2012
Agriculteurs exploitants	1,0
Artisans, commerçants, chefs d'entreprise	3,4
Cadres, professions intellectuelles supérieures	9,6
Professions intermédiaires	13,3
Employés	16,0
Ouvriers (y compris agricoles)	12,4
Inactifs ayant déjà travaillé	26,5
Autres sans activité professionnelle	17,7
Total	100,0

Champ : population des ménages de 15 ans ou plus, vivant en France métropolitaine. Résultats en moyenne annuelle. Source : INSEE, enquêtes Emploi.

Représentez graphiquement la répartition des PCS.

Solution

La représentation en secteur permet de bien faire apparaître l'importance relative de chaque catégorie.

Figure 11. Population de 15 ans et plus selon la catégorie socioprofessionnelle en 2012.

Les aires des disques seront proportionnelles aux effectifs de chacune des populations. C'est-à-dire :

$$\frac{\pi r_1^2}{\pi r_2^2} = \frac{A_1}{A_2} \text{ autrement dit } \frac{r_1}{r_2} = \sqrt{\frac{A_1}{A_2}}$$

La construction de plusieurs distributions circulaires est en général plus complexe que pour les diagrammes en colonnes.

Les représentations des variables quantitatives

Dans certains cas, la représentation des variables quantitatives utilise les représentations décrites ci-dessus. Il est possible de transformer une variable quantitative en variable qualitative, les valeurs de la variable ou les classes devenant alors les catégories de la variable qualitative. Les représentations graphiques préconisées pour les variables qualitatives sont alors applicables aux variables quantitatives transformées.
Deux représentations graphiques sont particulièrement adaptées aux variables quantitatives : le diagramme différentiel, pour les effectifs et les fréquences relatives et le diagramme cumulatif ou intégral pour les effectifs cumulés et fréquences cumulées.

Variable quantitative discrète

Le diagramme en bâtons est la représentation graphique différentielle des effectifs ou des fréquences d'une variable discrète. À chaque valeur (x_i) en

abscisse correspond un segment vertical de longueur proportionnelle soit à l'effectif (n_i), soit à la fréquence (f_i) de cette modalité. Ce graphique différentiel se distingue du graphique intégral ou cumulatif qui, lui, représente les fréquences cumulées. Le graphique intégral des fréquences cumulées représente la fonction cumulative ou fonction de répartition définie par $F(x_i) = F_i$, qui est une fonction étagée pour une variable discrète pour $x_i < x \leq x_{i+1}$.

Exercice : la répartition du nombre d'enfants par famille

Représentez graphiquement la répartition des familles selon le nombre d'enfants par un graphique différentiel et par un graphique cumulatif.

Tableau 31. Répartition des familles selon le nombre d'enfants.

Nombre d'enfants	0	1	2	3	4 et +	Total
Nombre de familles	7 000 000	3 600 000	3 300 000	1 300 000	500 000	15 700 000

Solution

Pour répondre à la question, il est nécessaire de construire le tableau statistique.

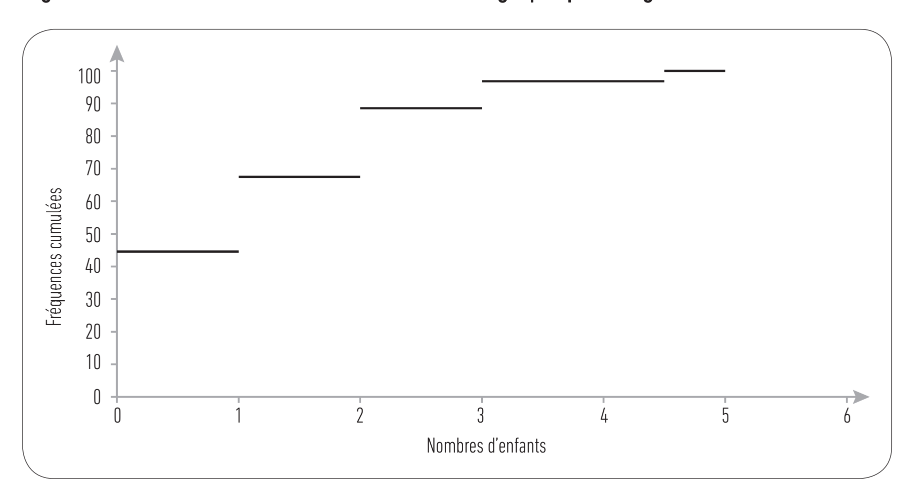
Tableau 32. Tableau statistique.

	Effectifs	**Fréquences relatives**	**Fréquences cumulées**
Nombre d'enfants	n_i	f_i	F_i
0	7 000 000	44,6	44,6
1	3 600 000	22,9	67,5
2	3 300 000	21,0	88,5
3	1 300 000	8,3	96,8
4,5	500 000	3,2	100,0
Ensemble	15 700 000	100,0	

Figure 12. Familles selon le nombre d'enfants (graphique différentiel).

Cette représentation indique clairement le caractère discret de la variable, le décalage de l'origine des abscisses est indispensable sinon le premier segment serait confondu avec l'axe des ordonnées.

Figure 13. Familles selon le nombre d'enfants (graphique intégral).



Ce graphique est discontinu pour chaque valeur entière de la variable.

Les variables continues

Le diagramme différentiel prend la forme d'un histogramme pour les séries classées. La forme du diagramme intégral est celle d'une courbe continue des fréquences cumulée.

L'histogramme est réservé aux séries groupées en classes. Pour visualiser l'importance relative des classes, on préfère les représenter par des surfaces en construisant un histogramme. L'histogramme est une représentation graphique de la distribution des effectifs ou des fréquences d'une variable

statistique continue ou considérée comme telle. À chaque classe de valeurs en abscisses, on fait correspondre un rectangle dont l'aire est proportionnelle à l'effectif de la classe (ou à la fréquence) : en abscisse l'amplitude de la classe, en ordonnée l'effectif (ou la fréquence) par unité d'amplitude. Soit une distribution $\{[b_i;b_{i+1}[\ ;\ n_i\}$ d'une variable statistique continue, pour chaque classe, l'histogramme associe un rectangle de largeur $a_i = b_{i+1} - b_i$ et de hauteur $h_i = \frac{f_i}{a_i}$.

Exemple : construction de l'histogramme de la répartition des surfaces agricoles utiles

La distribution des surfaces agricoles utiles d'une zone viticole est la suivante :

Tableau 34. Distribution des surfaces agricoles utiles.

Répartition en hectares	Nombre d'exploitations
[0 ; 5[	5 000
[5 ; 15[	15 000
[15 ; 25[	11 000
[25 ; 50[	20 000
[50 ; 100[	7 000
[100 ; 150]	2 000
Total	60 000

Solution

La première étape de toute représentation est la construction du tableau statistique.

Tableau 35. Répartition des SAU viticoles.

Classes (en ha)	n_i	a_i	f_i	$\frac{f_i}{a_i}$	F_i
[0 ; 5[	5 000	5	8,3	166,7	8,3
[5 ; 15[	15 000	10	25,0	250,0	33,3
[15 ; 25[	11 000	10	18,3	183,3	51,7
[25 ; 50[	20 000	25	33,3	133,3	85,0
[50 ; 100[	7 000	50	11,7	23,3	96,7
[100 ; 150]	2 000	50	3,3	6,7	100,0
Total	60 000		100,0		

Figure 14. Histogramme de la distribution.

Le polygone des fréquences lisse l'histogramme de façon à éliminer les ruptures qui dépendent du choix du découpage en classe. L'histogramme est fidèle au tableau de départ, il donne l'impression, l'illusion, qu'au sein de chaque classe, les valeurs sont régulièrement distribuées et qu'apparaissent des modifications brusques. La représentation ainsi obtenue apparaît plus réaliste ; la courbe de fréquences respecte la compensation des aires, théoriquement la surface située sous la courbe est identique à celle de l'histogramme. Cette représentation n'a pas de grande prétention scientifique, elle donne une image plus réaliste de la distribution.

Exemple : construction du polygone des fréquences de la répartition des surfaces agricoles utiles (SAU)

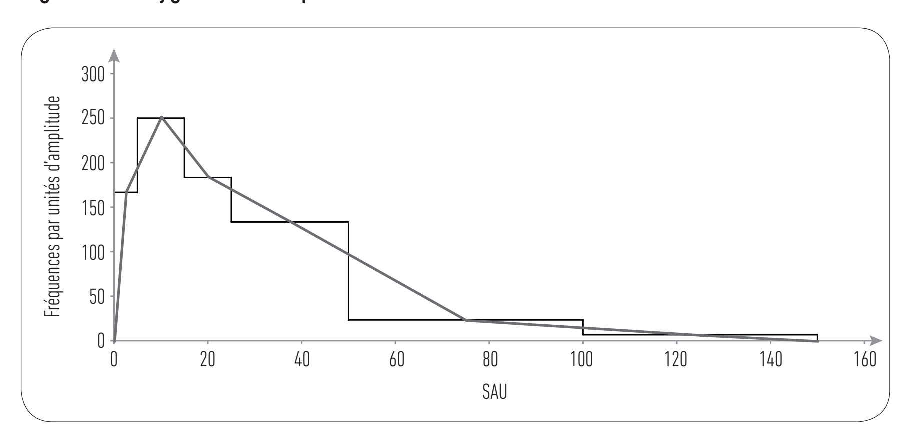
Ce polygone représente la distribution des fréquences. Les ordonnées sont les fréquences par unité d'amplitude, les abscisses sont des centres de classe. La courbe commence à la valeur minimale de la distribution, ici zéro, et s'achève à la valeur maximale de la dernière classe de SAU, ici 150. Le tableau suivant fournit les informations nécessaires pour construire le polygone. Construire le polygone des fréquences de la distribution des SAU.

Solution

Tableau 36. Tableau des données pour la construction du polygone des fréquences.

c_i	$\frac{f_i}{a_i}$
2,5	166,7
10,0	250,0
20,0	183,3
37,5	133,3
75,0	23,3
125,0	6,7

Figure 15. Polygone des fréquences.



Le polygone des fréquences donne une vision plus réaliste de la distribution en éliminant les ruptures entre les classes. Il permet également de percevoir la dissymétrie de la distribution.

La courbe cumulative des effectifs ou des fréquences (polygone des fréquences cumulées) représente graphiquement la fonction cumulative ou fonction de répartition définie par $F(x_i) = F_i$. La courbe cumulative des effectifs (ou des fréquences) s'obtient en joignant les points d'abscisse : la borne supérieure de la classe, et d'ordonnée : l'effectif cumulé croissant correspondant.

Exemple : construction de la courbe cumulative de la répartition des surfaces agricoles utiles

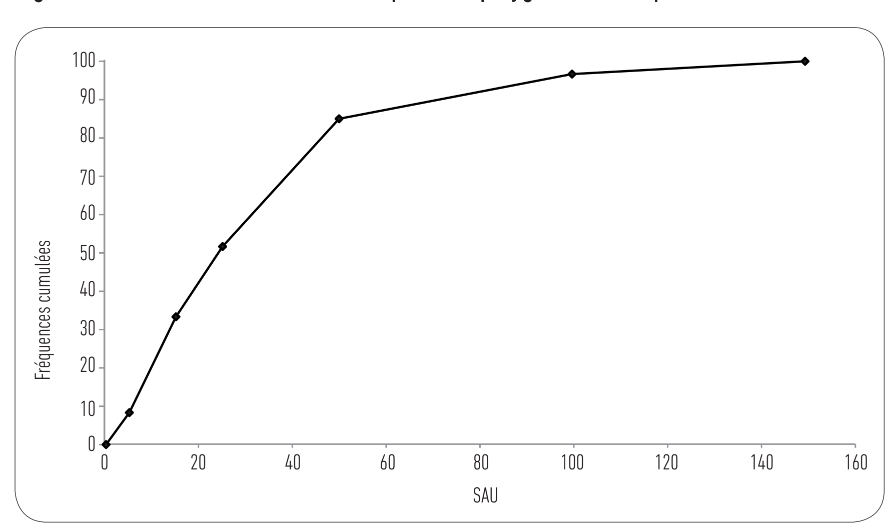
La courbe cumulative est une représentation des fréquences cumulées du tableau suivant.

Solution

Tableau 37. Fréquences cumulées de la répartition des SAU viticoles.

Classes (en ha)	F_i
[0 ; 5[	8,3
[5 ; 15[	33,3
[15 ; 25[	51,7
[25 ; 50[	85,0
[50 ; 100[	96,7
[100 ; 150]	100,0

Figure 16. Courbe cumulative des fréquences, polygone des fréquences cumulées.



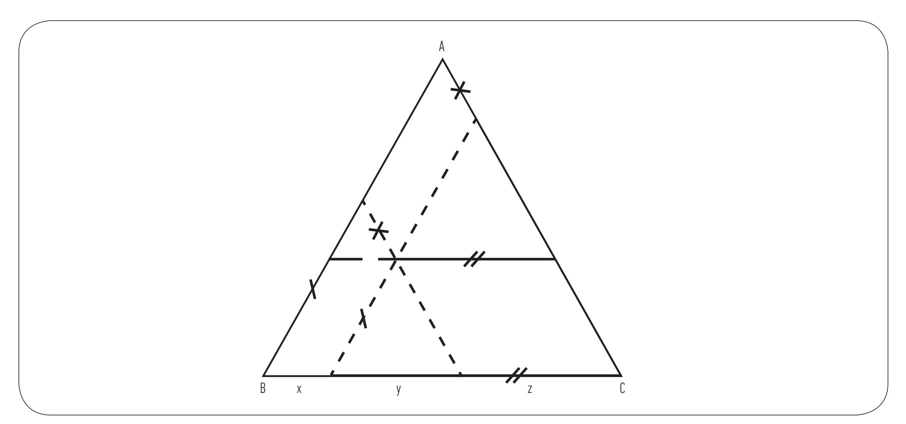
Cette courbe peut être considérée comme la fonction de répartition des SAU. Cette courbe des fréquences pourra être utilisée pour comparer la distribution réelle avec un modèle probabiliste connu.
En plus des représentations habituelles ci-dessus, il est apparu pertinent de présenter le diagramme triangulaire et le digramme polaire.

Le graphique triangulaire

Ce type de graphique est principalement utilisé pour représenter des séries statistiques constituées de trois variables dont la somme est constante ; le plus souvent, il s'agira de la répartition en pourcentage selon trois dimensions d'une grandeur variable.

L'utilisation du diagramme triangulaire repose sur une propriété du triangle équilatéral. D'un point M, intérieur au triangle, les droites tracées parallèlement aux côtés découpent des segments dont la somme est constante et égale à la longueur du côté.

Figure 17. Principe du diagramme triangulaire.



Si L est la longueur du côté du triangle alors a + b + c = L.

Exemple : les secteurs d'activité

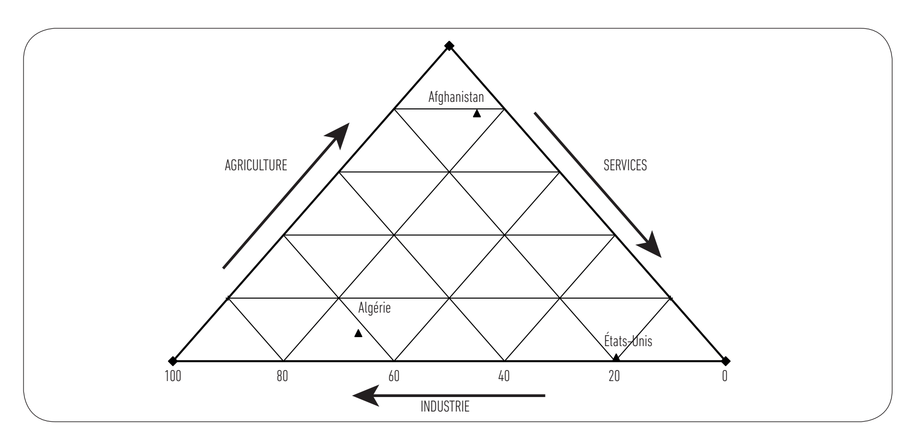
Tableau 38. Population active par secteur d'activité.

En %	Agriculture	Industrie	Services	Année
Afghanistan	78,6	5,7	15,7	2009
Algérie	8,9	62,0	29,1	2011
États-Unis	1,2	19,2	79,6	2011

Source : Banque mondiale

Solution

Figure 18. Population active par secteurs d'activité.

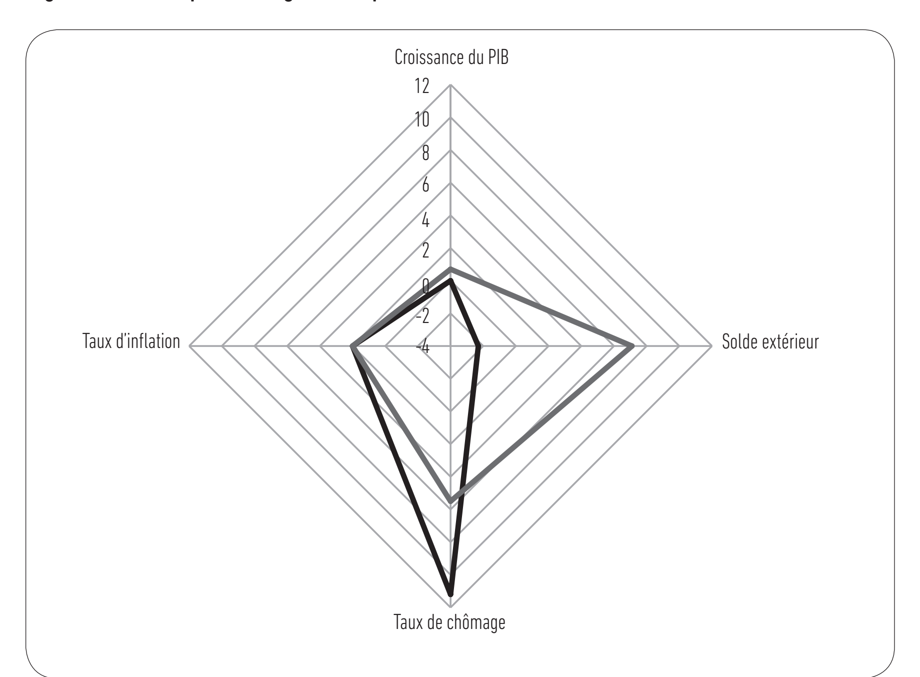


Outil utilisé : Comment faire un graphique ternaire ? Jacques Vaillé jacques.vaille@free.fr

Diagramme polaire

Pour terminer sur les différents types de représentation nous devons évoquer le diagramme polaire qui permet de visualiser un phénomène sur plusieurs axes. Dans un graphique à coordonnées cartésiennes, un point M est repéré par ses coordonnées (x et y) ; dans un graphique polaire, il l'est par l'angle θ (angle polaire) et la mesure algébrique ρ du vecteur $\vec{OM}$.

Figure 19. Principe du diagramme polaire.



Un exemple de ce type de graphique est connu sous le nom de carré magique de l'économiste britannique N. Kaldor résume la situation économique conjoncturelle d'un pays en retenant quatre objectifs de politique économique :

- la croissance économique : évaluée par le taux de croissance du PIB ;
- la situation de l'emploi : mesurée par le taux de chômage en pourcentage de la population active ;
- la stabilité des prix : mesurée par le taux d'inflation en pourcentage ;
- l'équilibre des comptes extérieurs : mesuré par le solde de la balance des paiements en pourcentage du PIB.

La situation conjoncturelle idéale est représentée un carré.

Les données relatives à la France et à l'Allemagne sont disponibles sur Eurostat. L'utilisation d'un graphique polaire permet une représentation révélatrice de la configuration conjoncturelle de chaque pays.

Exemple : représentation des situations conjoncturelles

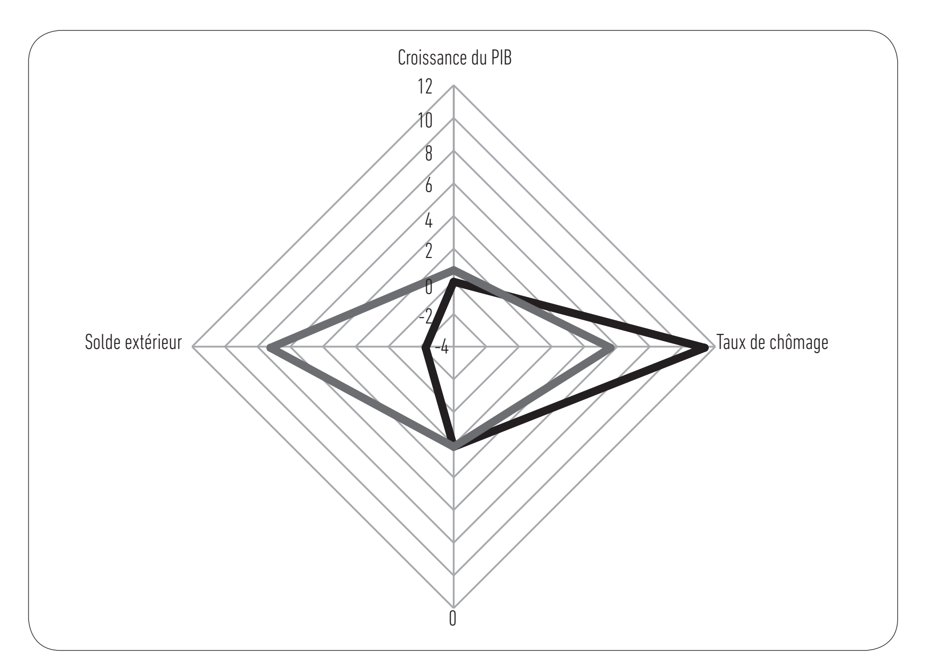
Tableau 39. Variations annuelles moyennes 2012.

	Croissance du PIB (%)	Solde extérieur (% du PIB)	Taux de chômage (%)	Taux d'inflation (%)
Allemagne	0,7	7,1	5,5	2,0
France	0,0	-2,3	11,2	2,0

Source : Eurostat

Pour représenter cette distribution à quatre dimensions, le graphique polaire à quatre axes est parfaitement adapté.

Figure 20. Le graphique polaire à quatre axes.



Les deux « carrés » sont assez dissemblables, traduisant des conjonctures économiques contrastées.
Ces premiers outils de traitement des données statistiques ne recourent qu'à des mathématiques élémentaires et à l'usage de logiciels faciles d'accès. Sans l'usage de ces outils, nombre de représentations graphiques seraient extrêmement complexes à produire. Cette facilité des traitements permet de se focaliser sur l'analyse. Toutefois dans ce premier chapitre, notre choix a été de ne pas développer les analyses pour s'en tenir à la définition et mise en œuvre des outils.

Chapitre 2

Les distributions à une dimension

Le terme de distribution doit être précisé ici. La distribution des valeurs d'un caractère définit la correspondance existant dans la population observée, entre une valeur – ou une classe de valeurs ou une modalité qualitative – et l'effectif ou la fréquence des individus affectés à cette valeur – voire d'une valeur de la classe ou de la modalité. Une distribution statistique peut être synthétisée par un seul chiffre, la *grandeur typique* de la distribution. Le graphique de la distribution met souvent en évidence une propension des valeurs à se concentrer au voisinage d'une valeur particulière. Une telle valeur symbolise en quelque sorte le cœur, le centre de gravité de la distribution ; elle en constitue la valeur essentielle, elle est dite *centrale*. Elle caractérise la distribution. Le calcul d'une caractéristique de tendance centrale modifie la perception de la distribution ; désormais, un seul chiffre représente un ensemble parfois fort vaste ; on peut ici évoquer l'exemple du salaire moyen en France, très éclairant à cet égard. L'intelligibilité de la situation en est simplifiée aux dépens de sa complexité.

Les valeurs centrales sont complétées par des caractéristiques de dispersion et des mesures de la concentration. Le statisticien britannique G. U. Yule a précisé les propriétés souhaitables pour un indicateur statistique :

- être défini de façon objective ;
- dépendre de toutes les observations ;
- avoir une signification concrète ;
- être simple à calculer ;
- se prêter aisément au calcul algébrique ;
- et être peu sensible aux fluctuations d'échantillonnage.

Aucun indicateur ne satisfait simultanément toutes les conditions dont certaines peuvent se révéler incompatibles. Le choix de l'indicateur sera fonction des données, des moyens de calculs disponibles et des objectifs présidant aux calculs.

Les tendances centrales

Une distribution statistique peut être synthétisée par un seul chiffre, par une valeur particulière, dite *centrale*.
Suivant le type de variable, nous verrons quelles caractéristiques il est possible de calculer. Le calcul des caractéristiques de tendance centrale est facilité par la construction de tableaux statistiques.

Tableau 1. Tableau statistique standard pour une variable discrète.

Valeurs de la variable	Effectifs	Fréquences	Fréquences cumulées		
x_i	n_i	f_i	F_i	$f_i x_i$	$f_i x_i^2$
x_1	n_1	$f_1 = \frac{n_1}{n}$	F_1	$f_1 x_1$	$f_1 x_1^2$
x_i	n_i	$f_i = \frac{n_i}{n}$	$F_i = \sum_{k=1}^{i} f_k$	$f_i x_i$	$f_i x_i^2$
x_m	n_m	$f_m = \frac{n_m}{n}$	$F_m = 1$	$f_m x_m$	$f_m x_m^2$
	$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$		$\sum_{i=1}^{m} f_i x_i$	$\sum_{i=1}^{m} f_i x_i^2$

Tableau 2. Tableau statistique standard pour une variable classée.

Centres des classes	Effectifs	Fréquences	Fréquences par unité d'amplitude	Fréquences cumulées		
c_i	n_i	f_i	$h_i = f_{i/a_i}$	F_i	$f_i c_i$	$f_i c_i^2$
c_1	n_1	$f_1 = \frac{n_1}{n}$	$h_1 = \frac{f_1}{a_1}$	F_1	$f_1 c_1$	$f_1 c_1^2$
c_i	n_i	$f_i = \frac{n_i}{n}$	$h_i = \frac{f_i}{a_i}$	$F_i = \sum_{k=1}^{i} f_k$	$f_i c_i$	$f_i c_i^2$
c_m	n_m	$f_m = \frac{n_m}{n}$	$h_m = \frac{f_m}{a_m}$	$F_m = 1$	$f_m c_m$	$f_m c_m^2$
	$n = \sum_{i=1}^{m} n_i$	$\sum_{i=1}^{m} f_i = 1$			$\sum_{i=1}^{m} f_i c_i$	$\sum_{i=1}^{m} f_i c_i^2$

Le mode

Le *mode* (noté M_O) d'une distribution est la valeur la plus fréquente dans la série. Il correspond à la valeur de la variable pour laquelle la fréquence est la plus élevée. C'est la valeur qui correspond au maximum du diagramme différentiel, le diagramme représentatif des effectifs ou des fréquences relatives. Dans le cas d'une variable non classée, caractère qualitatif ou variable discrète, la détermination du mode est immédiate, c'est la valeur ayant le plus grand effectif ou la plus grande fréquence.

Exemple d'un caractère qualitatif : les professions et catégories sociales

L'enquête « Emploi » de l'INSEE donne une répartition selon les PCS de la population active occupée.

Tableau 3. Population en emploi selon la catégorie socioprofessionnelle (en milliers).

	2003	2012
Agriculteurs exploitants	715,6	515,1
Artisans, commerçants, chefs d'entreprise	1 530,0	1 674,0
Cadres et professions intellectuelles supérieures	3 652,3	4 635,8
Professions intermédiaires	5 651,1	6 361,3
Employés	7 181,2	7 237,0
Ouvriers	5 897,9	5 356,9
Ensemble	24 677,5	25 754,3

Source : Insee, enquête Emploi 2012

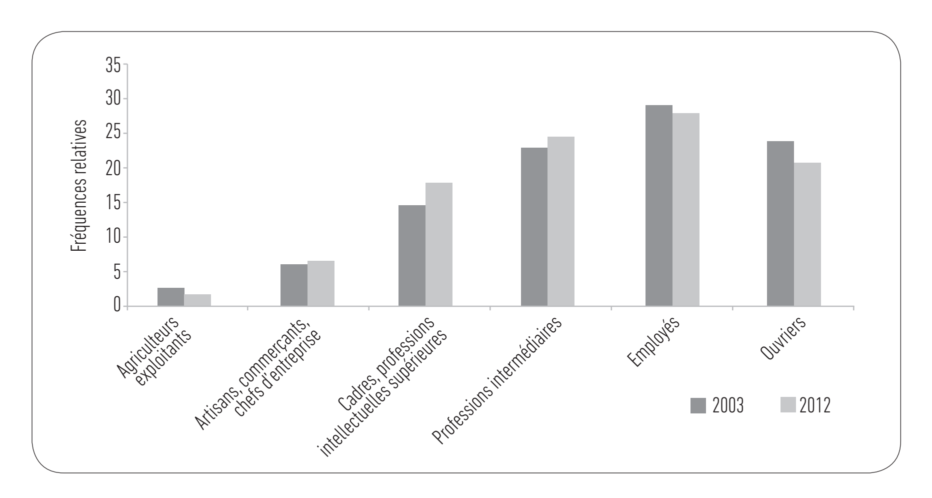
Le mode de cette distribution correspond à la catégorie « Employés » tant en 2003 qu'en 2012.
Le tableau des fréquences facilite l'analyse et la représentation graphique.

Tableau 4. Population en emploi selon la catégorie socioprofessionnelle en fréquences (en %).

	2003	2012
Agriculteurs exploitants	2,9	2,0
Artisans, commerçants, chefs d'entreprise	6,2	6,5
Cadres et professions intellectuelles supérieures	14,8	18,0
Professions intermédiaires	22,9	24,7
Employés	29,1	28,1
Ouvriers	23,9	20,8
Ensemble	100,0	100,0

Entre 2003 et 2012, la part des employés diminue alors que les professions intermédiaires progressent, dépassant les ouvriers.
Puisque l'objectif est de déterminer le mode d'un caractère qualitatif, la représentation graphique en colonnes est une des plus adaptées.

Figure 1. Les PCS des actifs occupés en 2003 et 2012.



Pour un caractère qualitatif, le mode est facile à déterminer. Le graphique permet de déterminer le mode, c'est la catégorie dont la colonne est la plus élevée, c'est-à-dire la catégorie « employés » pour l'année 2003 comme pour l'année 2012. Le tableau statistique nous indiquait que cette catégorie était la plus fréquente (29,1 % en 2003, 28,1 % en 2012).

Exemple de variable statistique discrète : le nombre d'enfants

Tableau 5. Répartition des familles selon le nombre d'enfants.

Nombre d'enfants	Effectifs des familles (en million)
0	7 492,3
1	3 615,8
2	3 255,3
3	1 267,0
4 et plus	465,4

Source : Insee, recensement 1999.

Le mode des familles ayant des enfants est « un enfant ». Pour l'ensemble des familles, le mode correspond aux familles n'ayant « aucun enfant ».

Exemple de variable statistique continue classée

Dans le cas d'une variable continue classée, la classe modale est celle dont la fréquence par unité d'amplitude notée $\frac{f_i}{a_i}$ ou h_i est la plus élevée. Il est ensuite possible de calculer la valeur du mode par la formule suivante :

$$M_o = b_i + \frac{\Delta_i}{\Delta_i + \Delta_{i+1}} \cdot a_i$$

avec M_o le mode, b_i limite inférieure de la classe modale (i), a_i l'amplitude de la classe modale, $\Delta_i = h_i - h_{i-1}$ différence entre la fréquence de la classe modale et la fréquence de la classe précédente dans la distribution, $\Delta_{i+1} = h_i - h_{i+1}$ différence entre la fréquence de la classe modale et la fréquence de la classe suivante dans la distribution. Pour illustrer ce cas, la série des SAU est éclairante.

Tableau 6. Répartition des SAU viticoles.

Classes (en ha)	a_i	c_i	n_i	f_i	$\frac{f_i}{a_i} \cdot 100$
[0 ; 5[	5	2,5	5 000	8,3	166,7
[5 ; 15[	10	10,0	15 000	25,0	250,0
[15 ; 25[	10	20,0	11 000	18,3	183,3
[25 ; 50[	25	37,5	20 000	33,3	133,3
[50 ; 100[	50	75,0	7 000	11,7	23,3
[100 ; 150]	50	125,0	2 000	3,3	6,7
			59 000	100,0	

Les $\frac{f_i}{a_i}$ du tableau sont calculés en multipliant par 100 le résultat de la division des fréquences par l'amplitude de façon à obtenir des données plus lisibles.

La classe modale se repère facilement sur l'histogramme, elle correspond à la classe ayant la fréquence par unité d'amplitude la plus élevée pour la distribution considérée, ici la seconde classe : [5 ; 15[.

L'histogramme permet de déterminer géométriquement le mode. Le mode est la valeur de l'abscisse correspondant à l'intersection des segments tracés sur le graphique ci-dessous.

Figure 2. Détermination géométrique du mode.

La détermination graphique permet ensuite de calculer la valeur du mode.

$$M_o = b_i + \frac{\Delta_i}{\Delta_i + \Delta_{i+1}} \cdot a_i = 5 + \frac{250 - 166,7}{250 - 166,7 + 250 - 183,3} \cdot 5 \cong 10,6 \text{ ha}$$

Le mode est facile à calculer et sa signification est immédiate, il se prête mal aux calculs algébriques ultérieurs.

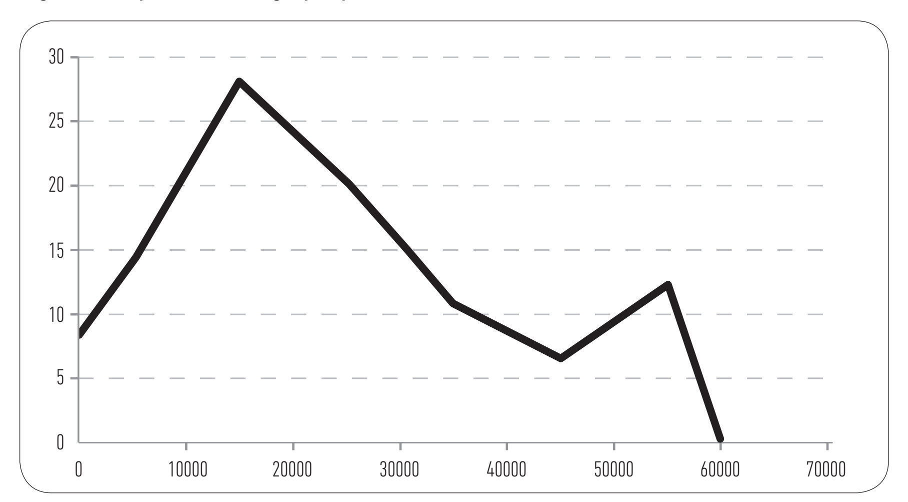
Les distributions statistiques les plus courantes n'ont qu'un seul mode (distribution unimodale), il arrive de rencontrer des distributions présentant plusieurs modes. Généralement, la présence de deux modes – les distributions bimodales – indique une population hétérogène composée de deux sous-populations. La distribution bimodale est obtenue par la superposition de deux distributions unimodales.

Tableau 7. Répartition des déclarations fiscales par tranche de revenus 2011 Belgique.

	Déclarations nulles	moins de 10 000 €	de 10 000 à 19 999 €	de 20 000 à 29 999 €	de 30 000 à 39 999 €	de 40 000 à 49 999 €	50 000 € et plus	60 000 €
Centre de classes	0	5 000	15 000	25 000	35 000	45 000	55 000	
Nombre	575 705	953 972	1 900 506	1 371 522	730 126	436 630	829 228	6 797 689
f_i	8,5	14,0	28,0	20,2	10,7	6,4	12,2	0

Source : DGSIE

Figure 3. Représentation graphique de cette distribution.



La représentation de cette distribution montre deux modes indiquant deux sous-populations ne disposant probablement pas des mêmes types de revenus.

La médiane

La médiane qui est notée M_e, d'une distribution ordonnée d'individus rangés selon la valeur croissante (respectivement décroissante) de la variable, est la valeur de la variable statistique qui partage la distribution en deux parties ayant des effectifs égaux.

La *médiane* est la valeur de la variable statistique telle que $F(M_e) = 0{,}5$, où F est la fonction de distribution représentée par les fréquences cumulées.

Il n'est pas toujours possible de déterminer la médiane dans le cas d'une variable discrète. Pour une variable continue, généralement classée, on détermine dans un premier temps une classe médiane. La classe i sera la classe médiane si $F_{i-1} < 0{,}5 \leq F_i$.

Puis, la médiane sera calculée par interpolation linéaire :

$$M_e = b_i + a_i \cdot \frac{50 - F_{i-1}}{F_i - F_{i-1}} = b_i + a_i \cdot \frac{50 - F_{i-1}}{f_i}$$

ou

$$M_e = b_{i+1} - a_i \cdot \frac{F_i - 50}{F_i - F_{i-1}} = b_{i+1} - a_i \cdot \frac{F_i - 50}{f_i}.$$

Exemple des SAU

Tableau 8. Répartition des SAU viticoles.

Classes (en ha)	a_i	f_i	F_i
[0 ; 5[	5	8,3	8,3
[5 ; 15[	10	25,0	33,3
[15 ; 25[	10	18,3	51,7
[25 ; 50[	25	33,3	85,0
[50 ; 100[	50	11,7	96,7
[100 ; 150]	50	3,3	100,0
		100,0	

La classe médiane est bien évidemment [15 ; 25[, la médiane est alors

$$M_e = b_i + a_i \cdot \frac{50 - F_{i-1}}{F_i - F_{i-1}} = 15 + 10 \cdot \frac{50 - 33{,}3}{18{,}3} \cong 24{,}1 \, ,$$

$$M_e = b_{i+1} - a_i \cdot \frac{F_i - 50}{f_i} = 25 - 10 \cdot \frac{51{,}7_i - 50}{18{,}3} = 24{,}1 \cdot$$

Il y a autant d'exploitation ayant une surface supérieure à 24,1 hectares que d'exploitation ayant moins de 24,1 hectares. Il est possible d'obtenir une évaluation de la médiane à partir de la courbe cumulative.

Figure 4. Détermination graphique de la médiane.

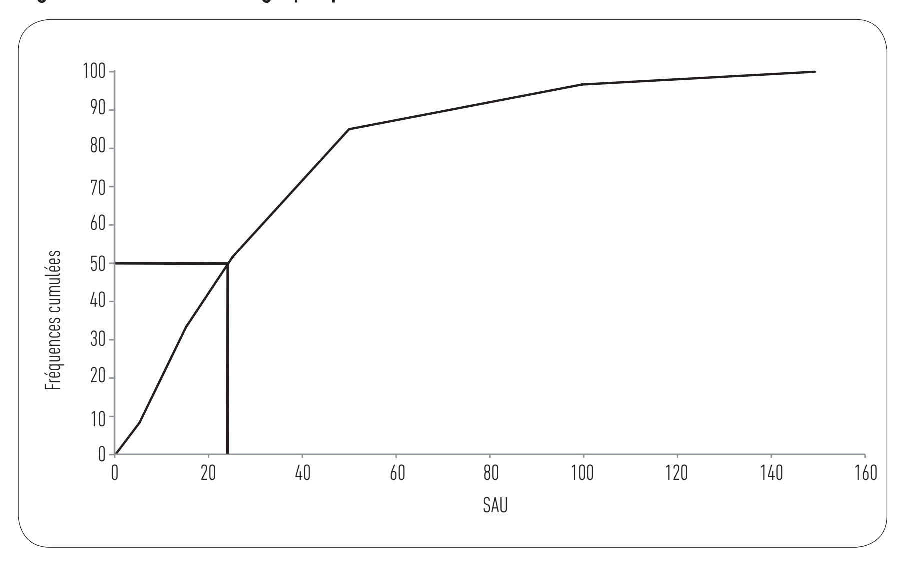
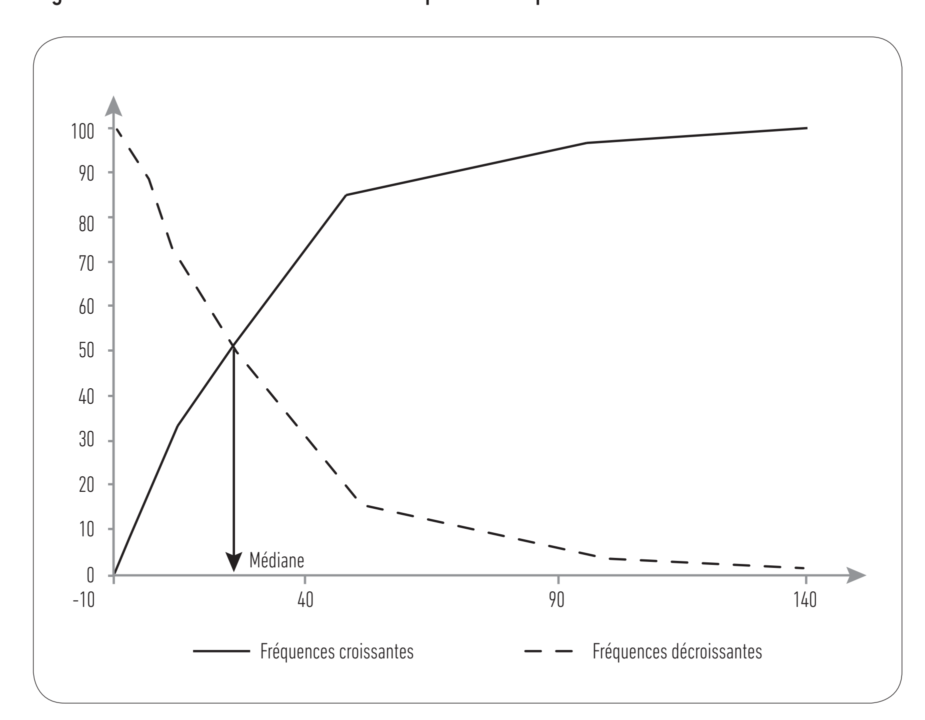


Figure 5. Détermination de la médiane par les fréquences.



La médiale

La *médiale* (notée M_l) est une caractéristique qui partage en deux moitiés égales la masse globale des valeurs. Pour une distribution salariale, le salaire médian est celui du salarié qui partage le nombre des salariés en deux groupes égaux ; le salaire médial est celui du salarié qui partage le total des salaires versés en deux groupes égaux.

Soit une distribution classée $\{(c_i, n_i)\ ; i \in [1,k]\}$, soit c_i les centres de classes, n_i l'effectif de la classe i. L'importance relative de la valeur de la classe i est :

$$f_i' = \frac{n_i c_i}{\sum_{i=1}^{k} n_i c_i} \text{ ou } f_i' = \frac{f_i c_i}{\sum_{i=1}^{k} f_i c_i} .$$

Nous noterons F'_i les fréquences cumulées de la série des valeurs. La médiale est alors définie par l'équation : $F'(M_l) = 0,5$. Le calcul de la médiane se fait par interpolation linéaire.

$$M_l = b_i + a_i \cdot \frac{50 - F'_{i-1}}{F'_i - F'_{i-1}} = b_i + a_i \cdot \frac{50 - F'_{i-1}}{f'_i} .$$

Exemple de médiale

Tableau 9. Tableau pour le calcul de la médiale de la distribution des SAU.

Classes (en ha)	a_i	c_i	f_i	f_ic_i	f'_i	F'_i
[0 ; 5[	5	2,5	8,3	20,8	0,7	0,7
[5 ; 15[	10	10,0	25,0	250,0	7,9	8,5
[15 ; 25[	10	20,0	18,3	366,7	11,5	20,1
[25 ; 50[	25	37,5	33,3	1 250,0	39,3	59,4
[50 ; 100[	50	75,0	11,7	875,0	27,5	86,9
[100 ; 150]	50	125,0	3,3	416,7	13,1	100,0
			100,0	3 179,2	100,0	

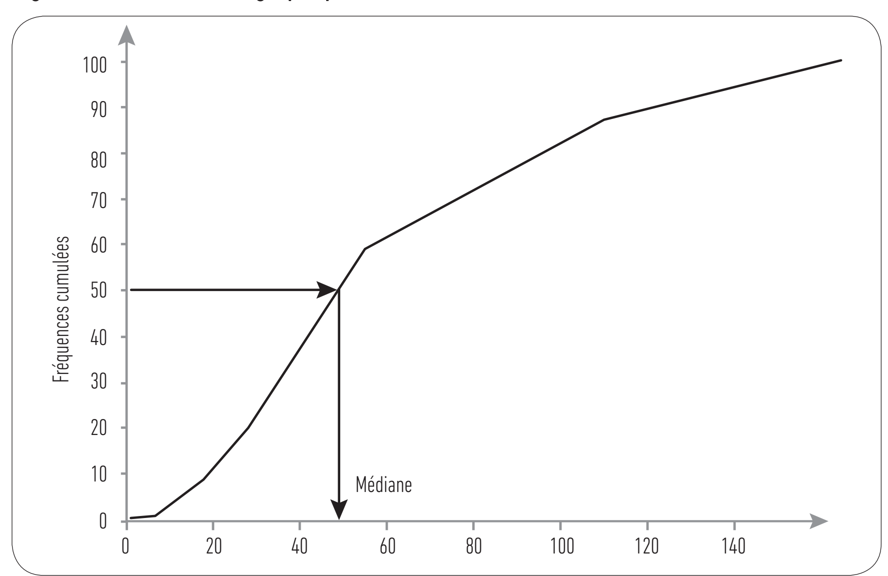
La classe médiale est celle renfermant $F'_i = 50$ donc la classe [25 ; 50[, la médiale est obtenue facilement par application de la formule de calcul.

$$M_l = b_i + a_i \cdot \frac{50 - F'_{i-1}}{F'_i - F'_{i-1}} = 25 + 25 \cdot \frac{50 - 20{,}1}{39{,}3} = 44$$

Les exploitations ayant moins de 44 hectares ont une surface identique à celle ayant plus de 44 hectares. La médiale est supérieure à la médiane, ce qui est une situation habituelle du fait des surfaces des exploitations les plus importantes dans la distribution.

$$\Delta M = M_l - M_e = 44 - 24{,}1 = 19{,}9$$

Figure 6. Détermination graphique de la médiale.



Les moyennes

Les caractéristiques de position, présentées ci-dessus, ne sont pas adaptées aux calculs algébriques, les moyennes répondent à cette contrainte. Une *moyenne* est une grandeur de tendance centrale calculée. Toutes les valeurs de la variable interviennent en fonction de leur importance relative, c'est-à-dire de leur *pondération*.

La moyenne arithmétique

La *moyenne arithmétique* (notée $\overline{x}$) est de loin la caractéristique de tendance centrale la plus usitée, celle dont on use et abuse sans toujours bien la comprendre. La moyenne arithmétique d'une variable statistique est la somme, pondérée par les fréquences, des valeurs. Quand la valeur de la variable n'apparaît qu'une fois, les pondérations sont alors toutes égales à 1.

Exemple pour une variable discrète

$$\overline{x} = \sum_{i=1}^{m} f_i x_i = \frac{1}{n}\sum_{i=1}^{m} n_i x_i$$

Tableau 10. Ménages selon le nombre de personnes (en milliers) 2010.

Nombre de personnes	x_i	n_i	f_i (en %)	$f_i x_i$
1	1	9 216,2	34,0	34,0
2	2	8 964,2	33,1	66,2
3	3	3 924,2	14,5	43,5
4	4	3 308,4	12,2	48,8
5	5	1 234,8	4,6	23,0
6 et plus	6,5	458,7	1,7	11,0
		27 106,5	100	226,6

Source : Ined : http://www.ined.fr

D'où la moyenne du nombre de personnes par logement, sous l'hypothèse que le nombre de personnes par logement pour les plus de 5 personnes soit de 6,5.

$$\overline{x} = \sum_{i=1}^{m} f_i x_i = \frac{226{,}6}{100} = 2{,}266 \cong 2{,}3$$

Exemple pour une variable classée

$$\overline{x} = \sum_{i=1}^{m} f_i c_i = \frac{1}{n} \sum_{i=1}^{m} n_i c_i \text{ avec } \sum_{i=1}^{m} f_i = 1 \text{ et } \sum_{i=1}^{m} \mathrm{n_i} = n$$

Tableau 11. Répartition des SAU viticoles.

Classes (en ha)	c_i	f_i	$f_i c_i$
[0 ; 5[	2,5	8,3	20,8
[5 ; 15[	10,0	25,0	250,0
[15 ; 25[	20,0	18,3	366,7
[25 ; 50[	37,5	33,3	1 250,0
[50 ; 100[	75,0	11,7	875,0
[100 ; 150]	125,0	3,3	416,7
		100,0	3 179,2

Si toutes les propriétés viticoles avaient toutes la même superficie, elle serait de 31,8 hectares. De façon équivalente, pour obtenir la surface globale, il suffit de multiplier le nombre d'exploitations par la superficie moyenne.

Quelques propriétés de la moyenne

Le résultat du calcul d'une moyenne, et cela s'applique à toutes les catégories de moyennes, dépend des hypothèses retenues pour les calculs. Les paragraphes ci-dessous illustrent quelques effets importants des variations suite au choix des hypothèses.

– La moyenne des différences à la moyenne

La moyenne des différences à la moyenne est nulle, en effet :

$$\sum_{i=1}^{k} f_i \left(x_i - \overline{x} \right) = 0 .$$

La démonstration est simple :

$$\sum_{i=1}^{k} f_i \left(x_i - \overline{x} \right) = \sum_{i=1}^{k} f_i x_i - \overline{x} \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} f_i x_i - \overline{x} = \overline{x} - \overline{x} = 0$$

car $\sum_{i=1}^{k} f_i = 1$.

Cette propriété donne, également, une définition de la moyenne arithmétique.

– Effet de regroupement

Pour une même distribution, selon le nombre de classes et le choix des regroupements[1], nous obtenons des salaires moyens différents. Un exemple sur la taille des conscrits illustre ce phénomène.

Tableau 12. Distribution des conscrits selon leur taille version 1.

Classe de taille en m	n_i	c_i	$n_i c_i$
[1,49 ; 1,75[	250	1,625	405
[1,75 ; 1,85[	50	1,80	90
Total	300		495

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i c_i = \frac{495}{300} = 1{,}65$$

La taille moyenne des conscrits est de 1,65 m. Il est possible de reclasser autrement cette même population à partir d'informations plus détaillées. Nous avons maintenant la distribution suivante.

Tableau 13. Distribution des conscrits selon leur taille version 2.

Classe de taille en m	n_i	c_i	$n_i c_i$
[1,50 ; 1,65[	30	1,625	48,75
[1,65 ; 1,75[	220	1,70	374
[1,75 ; 1,85[	50	1,80	90
Total	300		512,75

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i c_i = \frac{511{,}1}{300} = 1{,}71$$

La taille moyenne est de 1,71 m. Le conscrit moyen a grandi de 6 cm pour une même population. Ce résultat est dû à un simple effet de classement des données.

– Effet de bornes

Pour le calcul de la moyenne du nombre de personnes par logement, l'exercice précédent retenait une borne supérieure de 6,5 personnes. En modifiant la borne supérieure à 7,5 personnes, la moyenne est modifiée.

1. Il est bien clair que d'autres regroupements sont possibles en gardant le même nombre de classes.

Exemple :

Tableau 14. Ménages selon le nombre de personnes (en milliers) 2010.

Nombre de personnes	x_i	n_i	f_i (en %)	$f_i x_i$
1	1	9 216,2	34,0	34,0
2	2	8 964,2	33,1	66,1
3	3	3 924,2	14,5	43,4
4	4	3 308,4	12,2	48,8
5	5	1 234,8	4,6	22,8
6 et plus	7,5	458,7	1,7	12,7
		27 106,5	100	227,9

La moyenne devient 2,279 alors qu'avec une borne supérieure de 6,5 la moyenne est de 2,262. La différence paraît faible, cependant, si ces moyennes sont multipliées par le nombre de logements, soit 27 106,5 milliers, l'écart pour la population totale atteint 460 810,5 habitants, soit l'équivalent de Toulouse, la quatrième ville de France.
Les distributions sont généralement bornées vers le bas (revenu nul, salaire minimum), elles sont moins souvent limitées vers les valeurs élevées, la moyenne est sensible aux valeurs extrêmes et tend à s'accroître même pour des effectifs limités. Cette caractéristique la rend moins fiable que la médiane.

– La moyenne de sous-populations

La moyenne $\overline{x}$ d'une population P composée de deux sous-populations P_1 et P_2 s'exprime en fonction des moyennes des deux sous-populations.
Soit $\overline{x}_1$ la moyenne de la sous-population P_1 et de $\overline{x}_2$ la moyenne de la sous-population P_2.

$$\overline{x} = \sum_{i=1}^{k} f_i x_i \quad \overline{x}_1 = \sum_{i=1}^{k} f_{i1} x_i \quad \overline{x}_2 = \sum_{i=1}^{k} f_{i2} x_i$$

$$n_i = n_{i1} + n_{i2}$$

$$\frac{n_{i1}}{n_1} \cdot \frac{n_1}{n} + \frac{n_{i2}}{n_2} \cdot \frac{n_2}{n} = f_{i1} \cdot f_1 + f_{i2} \cdot f_2$$

$$\overline{x} = \sum_{i=1}^{k} \left(f_1 f_{i1} + f_2 f_{i2} \right) x_i = f_1 \sum_{i=1}^{k} f_{i1} x_i + \mathrm{f}_2 \sum_{i=1}^{k} f_{i2} x_i = f_1 \overline{x}_1 + f_2 \overline{x}_2 = \sum_{h=1}^{2} f_h x_h$$

$$\overline{x} = f_1 \overline{x}_1 + f_2 \overline{x}_2 = \sum_{h=1}^{2} f_h x_h$$

La moyenne $\overline{x}$ est la moyenne des moyennes des sous-populations P_1 et P_2.
La moyenne $\overline{x}$ d'une population P composée de sous populations P_h se généralise facilement fonction des moyennes des sous-populations. Soit $\overline{x}_h$ la moyenne, f_h l'importance de la sous-population P_h dans la population totale et $\overline{x}$ la moyenne de la population P.

$$\overline{x} = \sum_{i=1}^{k} f_i x_i \ ; \ \overline{x}_h = \sum_{i=1}^{k} f_{ih} x_i$$

$$\overline{x} = f_1 \overline{x}_1 + f_2 \overline{x}_2 + \cdots + f_k \overline{x}_k = \sum_{h=1}^{k} f_h x_h$$

La moyenne $\overline{x}$ est la moyenne des moyennes des sous-populations P_h.
Dans le calcul d'une moyenne, il est donc possible de remplacer des groupes de valeur par leur moyenne, puis de calculer la moyenne globale en utilisant les coefficients appropriés.

– L'effet de structure

La moyenne dépend des pondérations retenues pour le calcul. Elle est dite sensible aux effets de structure

Exemple :

Un exemple illustre ce phénomène, celui des salaires nets horaires moyens en 2010 qui sont connus par des enquêtes de l'INSEE. Les données pour quatre départements sont présentées dans le tableau suivant.

Tableau 15. Distribution des salaires nets horaires moyens 2010 (en €).

	Cadres	Professions intermédiaires	Employés	Ouvriers
Côte-d'Or	21,6	13,9	9,3	9,4
Isère	21,9	14,1	9,3	10,1
Paris	28,0	15,5	11,0	11,1
Vendée	20,7	13,2	8,6	9,6

Tableau 16. Distribution des salariés en fonction des PCS (2010).

	Cadres	**Professions intermédiaires**	**Employés**	**Ouvriers**
Côte-d'Or	15,1	28,1	30,5	26,3
Isère	19,8	28,4	28,1	23,7
Paris	35	26,9	27,8	10,3
Vendée	9,6	23,9	29,9	36,6

Le calcul des salaires moyens par régions fait apparaître des différences qu'il sera possible en partie d'expliquer par les structures différentes de la population active.

Tableau 17. Moyenne des salaires horaires pour quatre départements.

	x_i	f_i	$f_i x_i$
Côte-d'Or			
Cadres	21,6	15,1	3,3
Professions intermédiaires	13,9	28,1	3,9
employés	9,3	30,5	2,8
ouvriers	9,4	26,3	2,5
Ensemble		100,0	12,5
Isère			
Cadres	21,9	19,8	4,3
Professions intermédiaires	14,1	28,4	4,0
employés	9,3	28,1	2,6
ouvriers	10,1	23,7	2,4
Ensemble		100,0	13,3
Vendée			
Cadres	20,7	9,6	2,0
Professions intermédiaires	13,2	23,9	3,2
employés	8,6	29,9	2,6
ouvriers	9,6	36,6	3,5
Ensemble		100,0	11,2

Paris			
Cadres	28,0	35,0	9,8
Professions intermédiaires	15,5	26,9	4,2
employés	11,0	27,8	3,1
ouvriers	11,1	10,3	1,1
Ensemble		100,0	18,2

Les moyennes s'obtiennent immédiatement

Côte-d'Or : $\overline{x} = \sum_{i=1}^{k} f_i x_i = 12{,}5$ euros

Isère : $\overline{x} = \sum_{i=1}^{k} f_i x_i = 13{,}3$ euros

Paris : $\overline{x} = \sum_{i=1}^{k} f_i x_i = 18{,}2$ euros

Vendée : $\overline{x} = \sum_{i=1}^{k} f_i x_i = 11{,}2$ euros

La hiérarchie salariale obtenue est la suivante : Paris, Isère, Côte-d'Or, Vendée. Les taux de salaires horaires des quatre départements sont différents avec des taux bien plus élevés pour Paris. Les tableaux indiquent que les structures des emplois sont également différentes. Pour mettre en lumière les effets de structure, les calculs sont effectués en gardant les salaires horaires départementaux en utilisant les pondérations du département Paris, toute autre référence aurait été satisfaisante. Paris est souvent implicitement ou explicitement la référence en France du fait de la centralisation et de l'importance de la métropole dans l'activité nationale, d'un point de vue plus statistique, Paris est à la fois le département qui comprend le pourcentage le plus élevé de cadres et le plus faible pour les ouvriers ; en cela il est le plus atypique.

Tableau 18. Moyenne des salaires horaires pour trois départements avec les pondérations de Paris (f'_i).

	x_i	f_i	$f'_i x_i$
Côte-d'Or			
Cadres	21,6	35,0	7,6
Professions intermédiaires	13,9	26,9	3,7
employés	9,3	27,8	2,6
ouvriers	9,4	10,3	1,0
Ensemble		100,0	14,9
Isère			
Cadres	21,9	35,0	7,7
Professions intermédiaires	14,1	26,9	3,8
employés	9,3	27,8	2,6
ouvriers	10,1	10,3	1,0
Ensemble		100,0	15,1
Vendée			
Cadres	20,7	35,0	7,2
Professions intermédiaires	13,2	26,9	3,6
employés	8,6	27,8	2,4
ouvriers	9,6	10,3	1,0
Ensemble		100,0	14,2

Côte-d'Or : $\overline{x}' = \sum_{i=1}^{k} f'_i x_i = 14{,}9$ euros

Isère : $\overline{x}' = \sum_{i=1}^{k} f'_i x_i = 15{,}1$ euros

Paris : $\overline{x} = \sum_{i=1}^{k} f_i x_i = 18{,}2$ euros

Vendée : $\overline{x}' = \sum_{i=1}^{k} f'_i x_i = 14{,}2$ euros

Les écarts entre les départements diminuent sensiblement. Ils s'expliquent par les différences de taux de salaires horaires. Ce calcul permet d'éliminer l'effet de structure.

Tableau 19. Tableau des indicateurs des écarts.

Départements	Salaire net horaire moyen pondération locale (a)	Salaire net horaire moyen pondération Paris (b)	Rapport entre (b) et (a) $\frac{b}{a}$	Rapport (a) et Paris : $\frac{a}{Paris}$	Rapport entre (b) et Paris $\frac{b}{Paris}$
Côte-d'Or	12,5	14,9	119,0	68,7	81,7
Isère	13,3	15,1	113,0	73,5	83,0
Vendée	11,2	14,2	126,3	61,8	78,0

Avec la structure des salariés de Paris, le salaire moyen de la Côte-d'Or serait 19 % plus élevé que celui constaté, ce serait 13 % pour l'Isère et 26,3 % pour la Vendée. Les écarts relatifs entre les trois départements retenus et Paris sont moindres avec une structure identique des emplois.
La comparaison entre deux grandeurs moyennes doit comprendre une analyse de la structure des populations concernées.

– L'effet de sondage : redressement d'échantillon

Lors d'une enquête effectuée auprès d'un échantillon de 6 800 répondants, sur leur opinion concernant un nouvel aménagement sportif, les résultats ont été les suivants :

Tableau 20. Tableau des résultats.

Classe d'âge	Échantillon	Opinion favorable
de 15 à moins de 30 ans	1 000	0,85
de 30 à moins de 50 ans	2 000	0,54
50 ans et +	3 800	0,35
Total	6 800	

Tableau 21. Répartition par âges au sein de la population de référence.

Classe d'âge	Pourcentage
[15 ; 30[	0,25
[30 ; 50[	0,35
50 ans et +	0,40

Le pourcentage d'opinions favorables dans l'échantillon est la moyenne arithmétique des pourcentages d'opinions favorables.

Classe d'âge	Échantillon	Fréquences	Opinions favorables	
	n_i	f_i	x_i	$f_i x_i$
[15 ; 30[	1 000	0,15	0,85	0,1250
[30 ; 50[	2 000	0,29	0,54	0,1588
50 ans et +	3 800	0,56	0,35	0,1956
Total	6 800	1		0,4794

Dans l'échantillon, le taux d'opinions favorable est de 47,94 %, soit 48 %. La répartition par âge n'est pas identique dans l'échantillon et dans la population de référence. Il est donc nécessaire de redresser l'échantillon pour avoir une évaluation de la proportion d'opinions favorables dans la population de référence.
Le pourcentage d'opinions favorables dans la population est différent sous l'hypothèse que le pourcentage d'opinions favorables est une caractéristique de chaque classe d'âge.

Tableau 22. Écart entre échantillon théorique et échantillon empirique.

Classe d'âge	Structure de la population de référence	Échantillon théorique	Échantillon empirique	Coefficient de redressement
[15 ; 30[	0,25	1 700	1 000	1,70
[30 ; 50[	0,35	2 380	2 000	1,19
50 ans et +	0,4	2 720	3 800	0,72
Total	1	6 800	6 800	

Le coefficient de redressement est obtenu en divisant les effectifs de l'échantillon théorique par ceux de l'échantillon empirique.
Nous disposons de plus de répondants dans la dernière classe que dans les deux autres ; or les réponses sont plus favorables pour ces deux classes que pour la troisième. Nous devons donc supposer un biais qui minore le pourcentage d'opinions favorables dans la population de référence.
Nous allons détailler la méthode retenue en calculant ce que serait la répartition des effectifs.

Tableau 23. Comparaison des effectifs.

Classe d'âge	Effectifs empiriques favorables	Effectifs théoriques favorables
[15 ; 30[	850	1 445
[30 ; 50[	1 080	1 285
50 ans et +	1 330	952
Total	3 260	3 682

L'échantillon théorique des opinions favorables est obtenu en multipliant le nombre d'opinions favorables dans l'échantillon empirique par le coefficient de redressement, nous obtenons alors :

moyenne dans l'échantillon redressé $= \frac{3682}{6800} = 0,5414$.

Le pourcentage d'opinions favorables dans la population est donc de 54,14 % au lieu de 47,94 % comme dans l'échantillon empirique. Nous ne disposons pas des outils pour estimer si la différence entre les deux pourcentages est significative.
L'échantillon empirique diffère de l'échantillon théorique, il n'est pas fidèle à la structure de la population dans son ensemble. Sous réserve de régularité des opinions dans chaque catégorie de la population, il est possible de redresser l'échantillon et d'obtenir un résultat plus satisfaisant. Les résultats bruts conduisent à une majorité de défavorable alors qu'il est envisageable de supposer une légère majorité d'opinions favorables.

– L'effet Quetelet[2]

La connaissance de grandeurs moyennes pour une population ne permet de construire une population moyenne que si les relations entre les grandeurs sont linéaires. L'exemple du triangle moyen, dans un ensemble de triangles rectangles, illustre cette réflexion. Un premier triangle rectangle a pour côté 3,4 et 5, un second a respectivement des côtés de 5, 12 et 13. Quelles sont les dimensions des côtés du triangle moyen ? En prenant la moyenne de chaque côté, nous trouvons un triangle de côté 4, 8 et 9, ce qui ne correspond pas à des mesures d'un triangle rectangle ! Si la relation entre les variables n'est pas linéaire, le recours à la moyenne arithmétique n'est pas possible.

2. Pour reprendre le terme utilisé par Jean-Louis Boursin, *La statistique au quotidien*, Paris : Vuibert, 1992.

En conclusion, il est important de garder à l'esprit que les moyennes et en particulier la moyenne arithmétique résultent d'un calcul complexe dépendant d'un ensemble de paramètres. La comparaison de deux moyennes impose de vérifier que les hypothèses de calcul sont identiques.

La moyenne harmonique

La moyenne harmonique (notée H) est utilisée pour estimer la moyenne des inverses c'est en particulier le cas quand la grandeur à une dimension de vitesse.
La moyenne harmonique est définie par :

$$\frac{1}{H} = \frac{1}{n}\sum_{i=1}^{m}\frac{n_i}{x_i} = \sum_{i=1}^{m}\frac{f_i}{x_i} \quad \text{avec} \quad \sum_{i=1}^{k} f_i = 1 .$$

Exemple d'un calcul d'une vitesse moyenne

Sur le trajet de Paris à Grenoble, le TGV roule à 240 km/h sur 60 % du trajet, 150 km/h sur 25 % et à 120 km/h sur 15 % du trajet. À quelle vitesse moyenne parcourt-il l'ensemble du trajet ?
La vitesse moyenne entre Paris et Grenoble est obtenue par le rapport entre la distance parcourue et le temps mis pour la parcourir. La moyenne est donc une moyenne harmonique des vitesses. Nous pouvons le montrer simplement.

Soit x la distance entre Paris et Grenoble

$$x = v \cdot t$$

$$v = \frac{x}{t}$$

$$t = t_1 + t_2 + t_3$$

avec : $t = \frac{0,6x}{240} + \frac{0,25x}{150} + \frac{0,15x}{120}$

donc : $\frac{1}{v} = \frac{0,6}{240} + \frac{0,25}{150} + \frac{0,15}{120} = 0,00541667$.

Soit une vitesse moyenne de v = 184,6 km/h.
La vitesse moyenne est la moyenne harmonique des vitesses pondérées par la part de la distance parcourue. La vitesse moyenne est obtenue par application de la formule de définition de la moyenne harmonique :

$$\frac{1}{v} = \sum_{i=1}^{k}\frac{f_i}{v_i} .$$

Le calcul de la vitesse moyenne ne nécessite la connaissance ni de la distance à parcourir ni des temps correspondant à chaque fraction du trajet.

Il est facile de vérifier que la moyenne arithmétique est de 199,5 km/h ($0{,}6 \cdot 240 + 0{,}25 \cdot 150 + 0{,}15 \cdot 120 = 199{,}5$). Est-ce la vitesse moyenne du TGV ? Quel est le bon résultat ?

La distance Grenoble-Paris en TGV est de 579 kilomètres. Le tableau ci-dessous fournit les éléments de calcul de la vitesse moyenne sans recourir à la moyenne harmonique.

Tableau 24. Calculs de la vitesse moyenne.

Vitesses	Fréquences	Distances parcourues	Temps
v_i	f_i	$d_i = f_i \cdot 519$	$t_i = \frac{d_i}{v_i}$
240	0,6	347,4	1,4475
150	0,25	144,75	0,965
120	0,15	86,85	0,72375
	1	579	3,13625

La vitesse moyenne est alors $\overline{v} = \frac{d}{t} = \frac{579}{3{,}13625} \cong 184{,}6$

Le recours à la moyenne harmonique outre le fait qu'il est plus élégant est également plus rapide et représente la moindre source d'erreur.

Exemple : coût d'acquisition d'actions

Un épargnant a réalisé diverses opérations d'achat d'action d'une société cotée. Le cours étant de 9,9 € a-t-il intérêt à vendre ?

Tableau 25. Les achats de l'épargnant.

Montants des acquisitions (en €)	Valeur unitaire de l'action
v_i	x_i
30 000	5,6
11 000	8,8
22 000	10,3
17 000	8,2
18 000	7,8
5 000	11,0

Pour prendre sa décision, l'épargnant doit calculer la valeur moyenne d'acquisition des actions. Cette valeur est obtenue en calculant la moyenne harmonique des coûts d'acquisition des actions, en effet la valeur moyenne des actions acquises est le résultat du calcul suivant :

$$H = \frac{v}{n} = \frac{\sum_{i=1}^{6} v_i}{\sum_{i=1}^{6} n_i} = \frac{v}{\frac{v_1}{x_1} + \frac{v_2}{x_2} + \frac{v_3}{x_3} + \frac{v_4}{x_4} + \frac{v_5}{x_5} + \frac{v_6}{x_6}} = \frac{103000}{\frac{30000}{5,6} + \frac{11000}{8,8} + \frac{22000}{10,3} + \frac{17000}{8,2} + \frac{18000}{7,8} + \frac{5000}{11,0}}$$

Ce qui est très exactement le calcul d'une moyenne harmonique :

$$\frac{1}{H} = \frac{n}{v} = \frac{1}{103000}\left(\frac{30000}{5,6} + \frac{11000}{8,8} + \frac{22000}{10,3} + \frac{17000}{8,2} + \frac{18000}{7,8} + \frac{5000}{11,0}\right)$$

$$\frac{1}{H} = \frac{1}{v}\sum_{i=1}^{m} \frac{n_i}{v_i} = \frac{1}{103000} \cdot 13578,5 = 0,131829842$$

$$H = \frac{1}{0,131829842} \cong 7,6 \,.$$

En vendant au taux de 9,9, l'épargnant réalise un bénéfice de 9,9-7,6 = 2,3 € par action. Le recours à la moyenne harmonique évite d'avoir à déterminer le nombre des actions acquises sachant que dans le prix d'achat sont compris divers coûts annexes.

Un autre calcul est envisageable en calculant le nombre d'actions théoriques à chaque opération. Les résultats sont identiques.

Tableau 26. Une méthode alternative.

Montants des acquisitions	Valeur unitaire de l'action	nombre d'actions
v_i	x_i	$n_i = \frac{v_i}{x_i}$
30000	5,6	5 357,1
11000	8,8	1 250,0
22000	10,3	2 135,9
17000	8,2	2 073,2
18000	7,8	2 307,7
5000	11	454,5
103000		13 578,5

Le coût d'acquisition d'une action est obtenu par le rapport entre la totalité du coût des acquisitions sur le nombre théorique d'actions.

$$\overline{x} = \frac{\sum_{i=1}^{6} v_i}{\sum_{i=1}^{6} n_i} = \frac{103000}{13578,5} \cong 7,6$$

La moyenne harmonique est également utilisée pour le calcul des cours moyens des devises.

La moyenne géométrique

La moyenne géométrique est utilisée quand on étudie les variations relatives, en particulier les accroissements. La moyenne géométrique est utilisée pour calculer le multiplicateur moyen d'une grandeur.

Soit Y une grandeur qui croît successivement de r_1, r_2,…,r_i

$$Y_i = Y_0 (1+r_1)(1+r_2)\cdots(1+r_{i-1})(1+r_i) = Y_0 \ (1+r)^i$$

Le multiplicateur moyen 1+r est obtenu en calculant $(1+r) = \sqrt[i]{\frac{Y_i}{Y_0}}$, qui est la moyenne géométrique des multiplicateurs.

La formule générale *moyenne géométrique* (notée G) est la suivante :

$$G = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}$$

avec $\sum_{i=1}^{k} n_i = n$

ou sous forme logarithmique $Log\ \mathrm{G} = \frac{1}{n}\sum_{i=1}^{k} n_i Log\ x_i = \sum_{i=1}^{k} f_i Log\ x_i$

Exemple de calculs d'un taux de croissance moyen

Les variations annuelles du PIB en valeur d'un pays donné sont les suivantes par rapport à l'année précédente.

Tableau 27. Les taux annuels de croissance.

Années	1	2	3	4	5
Taux de croissance (en %)	10,0	8,5	7	5,6	8,0

- Calculer le taux de croissance annuel moyen en valeur de l'année 0 à l'année 5 ?
- Calculer le taux de croissance annuel moyen en valeur de l'année 1 à l'année 5 ?

Solution

- 1. Taux de croissance moyen entre l'année 0 et l'année 5

Nous ne pouvons pas obtenir directement le taux de croissance, nous devons tout d'abord calculer le multiplicateur moyen (1+r) qui est la moyenne géométrique des multiplicateurs annuels.

$$(1+r)=\sqrt[5]{\prod_{i=1}^{5}(1+r_i)}$$

Le tableau ci-dessous fournit la série des multiplicateurs :

Tableau 28. Tableau des multiplicateurs.

Années	1	2	3	4	5	5/0	5/1
Multiplicateurs	1,10	1,085	1,07	1,056	1,08	1,456444282	1,324040256

Entre l'année 1 et l'année 5, le multiplicateur moyen est :
$1+r=\sqrt[5]{1,456444282}\cong 1,0781$.
Le taux de croissance moyen est de 7,8 %.

– 2. Taux de croissance moyen entre l'année 1 et l'année 5

Le calcul est analogue sur quatre ans :
$1+r=\sqrt[4]{1,324040256}\cong 1,0727$.
Le multiplicateur est d'environ 1,073, ce qui correspond à un taux de croissance moyen de 7,3 %.

La moyenne quadratique

La moyenne quadratique (notée Q^2) est utilisée pour calculer des moyennes de carrés.

$$Q^2=\sum_{i=1}^{m} f_i x_i^2=\frac{1}{n}\sum_{i=1}^{m} n_i x_i^2 \text{ avec } \sum_{i=1}^{k} f_i=1$$

Exemple : Taille du côté moyen de la parcelle moyenne

Nous disposons d'un certain nombre de parcelles forestières de forme carrée. En vue d'améliorer l'exploitation des bois, les responsables veulent pouvoir disposer de la valeur du côté moyen de la parcelle moyenne.

Tableau 29. Distribution des parcelles.

Nombre de parcelles	Côté de chaque parcelle
45	9
25	11
15	12
10	17

– 1. Quelle est la moyenne arithmétique des côtés de cet ensemble de parcelles ?
– 2. Quelle est la surface totale des parcelles ?
– 3. Quelle est la moyenne quadratique ?

Il est possible de calculer la moyenne arithmétique des côtés.

La moyenne arithmétique

Tableau 30. Calcul du côté moyen des parcelles.

Nombre de parcelles	Côté des parcelles	
n_i	x_i	$n_i x_i$
45	9	405
25	11	275
15	12	180
10	17	170
95		1 030

La moyenne arithmétique des côtés des parcelles est donc :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i . x_i = \frac{1030}{95} = 10,842$$

Calcul de la surface

La surface obtenue est de $10,842^2$ x 95 = 11 167,37.

Or la surface réelle est de 11 720 comme le montre le calcul dans le tableau suivant soit une erreur de 552,6 m^2 ou 4,7 % de la surface totale.

Tableau 31. Calcul de la surface totale.

Côté des parcelles	Nombre de parcelles	Surfaces
9	45	3 645
11	25	3 025
12	15	2 160
17	10	2 890
		11 720

Le calcul de la moyenne arithmétique des côtés fournit une information erronée.

Calcul de la moyenne quadratique

Tableau 32. Calcul de la moyenne quadratique.

Côté des parcelles	Nombre de parcelles	
n_i	n_i	$n_i x_i^2$
9	45	3 645
11	25	3 025
12	15	2 160
17	10	2 890
	95	11 720

La moyenne quadratique est donc :

$$Q^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 = \frac{11720}{95} = 123,368 \text{ .}$$

La moyenne quadratique est celle de la surface moyenne des parcelles, la surface totale est : 123,368 x 95 = 11 719,96. L'écart avec la surface d'ensemble est de 0,04 m^2. Cet écart résulte des arrondis de calcul.

Le côté moyen est donc $Q = \sqrt{123,368421} \cong 11,1071$, soit environ 11,11 m. La moyenne quadratique des côtés des parcelles, donc la surface moyenne des parcelles donne une information correcte.

Une généralisation de la moyenne : la notion de Φ moyenne

Soit Φ(x) une fonction monotone c'est-à-dire une fonction toujours croissante ou toujours décroissante sur l'intervalle des calculs de la variable x, le nombre M sera la Φ-moyenne telle que :

$$\Phi(M) = \sum_{i=1}^{k} f_i \Phi(x_i) \text{ .}$$

Elle correspond à la définition générale de la moyenne.

Exemples :

$$\Phi(x) = \frac{1}{x}$$

$\frac{1}{M} = \sum_{i=1}^{k} f_i \frac{1}{x_i} w$, on reconnaît la moyenne harmonique.

$$\Phi(x) = Lnx$$

$Ln\ G = \sum_{i=1}^{k} f_i Ln\ x_i$, on reconnaît la moyenne géométrique.

$\Phi(x) = x$

$\overline{x} = \sum_{i=1}^{k} f_i x_i$, on reconnaît la moyenne arithmétique.

$\Phi(x) = x^2$

$Q^2 = \sum_{i=1}^{k} f_i x_i^2$, donne la moyenne quadratique.

Il est tout à fait possible aussi de construire de nouvelles moyennes :

$\Phi(x) = x^3$,

$M_3^3 = \sum_{i=1}^{k} f_i x_i^3$.

Les moyennes d'une même distribution respectent les inégalités suivantes :

$H < G < \overline{x} < Q$.

Une tendance centrale n'est pas toujours une caractéristique significative. En effet, si la recherche d'une tendance centrale est pertinente encore faut-il choisir la caractéristique significative. Les calculs de caractéristiques de tendance centrale sont souvent utilisés pour comparer des populations, il faut dans ce cas que les populations soient directement comparables. La comparaison des revenus monétaires moyens d'une population agricole dont l'autoconsommation est importante et d'une population urbaine salariée n'a pas qu'une pertinence discutable : la moyenne arithmétique est affectée par les valeurs extrêmes de la distribution, elles peuvent être aberrantes, d'où la préférence parfois accordée à la médiane.

La dispersion

Choisir un seul nombre pour résumer toute une distribution donne une information incomplète. Il est donc nécessaire de disposer d'informations sur la dispersion des valeurs autour de la caractéristique retenue. Les caractéristiques de dispersion sont absolues ou relatives, d'autres estiment la concentration. Suivant le type de variable, nous verrons quelles caractéristiques il est possible de calculer.

L'étendue

L'étendue *e*, autrement appelée l'intervalle de variation ou l'amplitude de la distribution est la plus simple des caractéristiques de dispersion. C'est la différence entre la plus grande et la plus petite valeur observée. L'importance relative de chaque valeur n'intervient pas dans le calcul de l'étendue qui

n'implique que les valeurs des extrémités du segment de la distribution, ce qui conduit à des fluctuations considérables. Si la dispersion est grande et les valeurs extrêmes peu nombreuses, l'étendue accentue la dispersion. Sa signification est claire, son calcul rapide, cependant c'est une caractéristique de dispersion peu significative, car trop dépendante de la distribution utilisée. L'étendue de la distribution des exploitations viticoles est de 150 ha. (150-0)

Les quantiles

Les quantiles partagent une distribution ordonnée en *n* parties ayant toutes le même effectif ; la médiane est un quantile qui partage la distribution en deux. Plusieurs types de partitions sont utilisés, les plus courantes sont : la médiane (division en deux), les quartiles (division en quatre), les quintiles (division en cinq), les déciles (division en dix). Les centiles (division en 100) et les terciles (divisions en trois) sont aussi employés, mais plus rarement. Les institutions internationales privilégient les déciles et les quintiles plus explicatifs que les quartiles et plus faciles à calculer, car demandant moins de données précises que les centiles.

Nous allons d'abord étudier les quartiles avant de nous pencher sur d'autres types de quantiles.

Les quartiles Q_1, Q_2, Q_3 partagent la série en quatre parties d'effectifs égaux comprenant chacun 25 % des effectifs. 25 % des données sont inférieures à Q_1, 25 % des données sont supérieures à Q_3, il est important de rappeler que le deuxième quartile est la médiane.

Le calcul d'un quartile se réalise par interpolation linéaire comme pour la médiane. Les quartiles sont solutions des trois équations $F(Q_1) = 0{,}25$; $F(Q_2) = 0{,}50$; $F(Q_3) = 0{,}75$.

Si la classe i est la classe contenant le premier quartile :

$$Q_1 = b_i + a_i \cdot \frac{25 - F_{i-1}}{F_i + F_{i-1}} \cdot$$

Si la classe i est la classe contenant le troisième quartile :

$$Q_3 = b_i + a_i \cdot \frac{75 - F_{i-1}}{F_i + F_{i-1}} \cdot$$

En ce qui concerne les quintiles, notés par V_1, V_2, V_3 et V_4, ils divisent une série statistique ordonnée en cinq groupes égaux comprenant chacun 20 % des données de la série : 20 % des données sont inférieures à V_1, 20 % des données sont supérieures à V_4.

À l'identique du quartile, la formule de calcul d'un quintile utilise l'interpolation linéaire, les quintiles extrêmes sont solutions des équations $F(V_1) = 0{,}20 \;\; ; \; F(\mathrm{v}_4) = 0{,}80$.
Si la classe i est la classe contenant le premier quartile :

$$V_1 = b_i + a_i \cdot \frac{20 - F_{i-1}}{F_i + F_{i-1}} \cdot$$

Si la classe i est la classe contenant le troisième quartile :

$$V_4 = b_i + a_i \cdot \frac{80 - F_{i-1}}{F_i + F_{i-1}} \cdot$$

Les déciles partagent la série en dix parties d'effectifs égaux

$$D_1 = b_i + a_i \cdot \frac{10 - F_{i-1}}{F_i + F_{i-1}}$$

$$D_9 = b_i + a_i \cdot \frac{90 - F_{i-1}}{F_i + F_{i-1}}$$

Les centiles partagent la distribution en 100 parties d'effectifs égaux. Les calculs sont analogues aux précédents.

Les intervalles interquantiles et rapports interquantiles

L'intervalle interquantile est le segment compris entre le premier et le dernier quantile, $[\, Qa_1, Qa_n]$ soit $[Q_1, Q_3]$ pour les quartiles, $[D_1, D_9]$ pour les déciles, $[V_1, V_4]$ pour les quintiles. Les intervalles interquartile et interdécile sont les plus courants. Ils donnent une première image de la dispersion des valeurs par rapport à l'étendue ; ils éliminent les valeurs extrêmes et offrent de ce fait une information plus significative.

L'écart interquantile est obtenu en faisant la différence entre le dernier et le premier quantiles. Il mesure l'étendue de l'intervalle inter quantile $I_{Qa} = Qa_n$ - Qa_1 soit $I_Q = Q_3 - Q_1$ pour les quartiles, $I_V = V_4 - V_1$ pour les quintiles, $I_D = D_9 - D_1$ pour les déciles. Il élimine l'effet des valeurs extrêmes ou aberrantes, ne se calcule que sur deux valeurs et il est facile à calculer et à interpréter. Attention cependant car il ne se prête pas très bien au calcul algébrique. Il contient un pourcentage des observations en fonction des quantiles retenus (50 % pour les quartiles, 60 % pour les quintiles, 80 % pour les déciles).

Le rapport interquantile : $IQ(x) = \frac{Q_n}{Q_1}$ est un nombre sans dimension qui mesure le rapport entre le dernier et le premier quantiles.

Prenons l'exemple des quartiles. Le rapport interquartile est le rapport des quartiles $IQ = \frac{Q_3}{Q_1}$, il fournit une mesure relative des écarts entre les 25 % de la distribution ayant les valeurs les plus basses et les 25 % de la distribution disposant des valeurs de la variable les plus élevées. Le rapport inter quintile $IV = \frac{V_4}{V_1}$ donne une mesure relative des écarts entre, par exemple, les 20 % de la population disposant des revenus les plus élevés et les 20 % de la population ayant les revenus les plus bas. Le rapport interdécile $ID = \frac{D_9}{D_1}$ fournit une mesure du rapport relatif entre les 10 % des valeurs plus élevés et les 10 % les plus faibles

Tableau 33. Distribution des salaires mensuels nets en 2012 dans la fonction publique d'État.

	Salaires 2012 (€)	Évolution 2011-2012 (% en € constants)
D1	1 484	- 0,3
D2	1 774	- 0,6
D3	1 945	- 1,1
D4	2 100	- 1,0
D5 (médiane)	2 259	- 1,1
D6	2 436	- 1,1
D7	2 664	- 1,0
D8	2 995	- 1,0
D9	3 571	- 1,0
D9/D1	2,4	0,0 point
Moyenne	2 465	

Source : Insee, SIASP.

L'écart interquantile fournit une information en valeur absolue de l'importance de l'inégalité, le rapport donne une indication relative de la disparité. Il s'agit d'indicateurs simples, faciles à comprendre donnant une première information. Les rapports sont en général plus utilisés que les écarts, car ils sont indépendants des unités de mesure en particulier de la monnaie et facilitent de la sorte les comparaisons internationales.

Le *coefficient de dispersion* utilise les écarts interquantiles en référence à la médiane de façon à obtenir un indicateur de dispersion relative sans dimension comme le rapport interquantile.

$$CDis_{Qn} = \frac{Qa_n \ - \ Qa_1}{M_e}$$

$$CDis_D = \frac{D_9 \text{ - } D_1}{M_e}$$

$$Cdis_V = \frac{V_4 - V_1}{M_e}$$

$$Cdis_Q = \frac{Q_3 - Q_1}{M_e}$$

Exemple

Tableau 34. Répartition des SAU viticoles.

Classes (en ha)	a_i	f_i	F_i
[0; 5[	5	8,3	8,3
[5; 15[	10	25,0	33,3
[15; 25[	10	18,3	51,7
[25; 50[	25	33,3	85,0
[50 ; 100[	50	11,7	96,7
[100 ; 150]	50	3,3	100,0
		100,0	

Nous obtenons par application des formules les indicateurs liés :
– aux quartiles

$$M_e = b_i + a_i \cdot \frac{50 - F_{i-1}}{F_i - F_{i-1}} = 15 + 10 \cdot \frac{50 - 33,3}{18,3} \cong 24,1 \text{ ha}$$

$$Q_1 = b_i + a_i \cdot \frac{25 - F_{i-1}}{F_i + F_{i-1}} = 5 + 10 \cdot \frac{25 - 8,3}{25} \cong 11,7 \text{ ha}$$

$$Q_3 = b_i + a_i \cdot \frac{75 - F_{i-1}}{F_i + F_{i-1}} = 25 + 25 \cdot \frac{75 - 51,7}{33,3} \cong 42,5 \text{ ha}$$

$$\text{I}_\text{Q} = \text{Q}_3 \text{ - } \text{Q}_1 = 41,5 - 11,7 = 29,8$$

$$IQ = \frac{Q_3}{Q_1} = \frac{41,5}{11,7} \cong 3,6$$

$$Cdis_Q = \frac{Q_3 - Q_1}{M_e} = \frac{29,8}{24,1} \cong 1,2$$

– aux déciles

$$D_1 = b_i + a_i \cdot \frac{10 - F_{i-1}}{F_i + F_{i-1}} = 10 + 5 \cdot \frac{10 - 8,3}{25} \cong 5,7 \text{ ha}$$

$$D_9 = b_i + a_i \cdot \frac{90 - F_{i-1}}{F_i + F_{i-1}} = 50 + 50 \cdot \frac{90 - 85}{23,3} \cong 57,5 \text{ ha}$$

$$I_D = D_9 - D_1 = 57,5 - 5,7 = 48,8$$

$$ID = \frac{D_9}{D_1} = \frac{57,5}{5,7} \cong 10,1$$

$$Cdis_D = \frac{D_9 - D_1}{M_e} = \frac{48,8}{24,1} \cong 2,0$$

– aux quintiles

$$V_1 = b_i + a_i \cdot \frac{20 - F_{i-1}}{F_i + F_{i-1}} = 5 + 10 \cdot \frac{20 - 8,3}{25} \cong 9,7 \text{ ha}$$

$$V_4 = b_i + a_i \cdot \frac{80 - F_{i-1}}{F_i + F_{i-1}} = 25 + 25 \cdot \frac{80 - 51,7}{33,3} \cong 46,3 \text{ ha}$$

$$I_V = V_4 - V_1 = 46,3 - 9,7 = 36,6 \text{ ha}$$

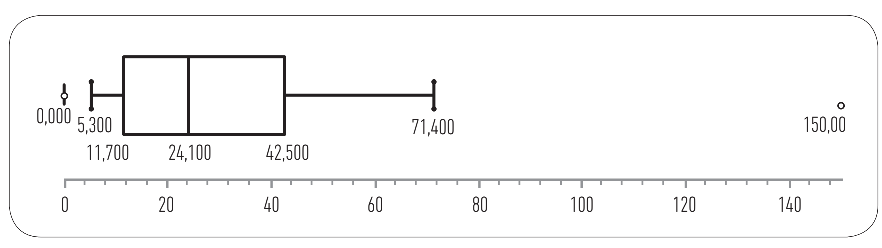
$$IV = \frac{V_4}{V_1} = \frac{46,3}{9,7} = 4,8$$

$$Cdis_V = \frac{V_4 - V_1}{M_e} = \frac{36,6}{24,1} \cong 1,5$$

Une représentation des quantiles : Le diagramme en boîte

Le diagramme en boîte – autrement appelé boîte à moustaches, *box plot* ou *box-and-whisker plot* – donne une représentation très simple de la distribution. Elle consiste en une boîte rectangulaire, dont les deux extrémités sont les quartiles. Ces extrémités se prolongent en segments dont les valeurs extrêmes sont les déciles. On représente aussi la médiane par un trait dans la boîte ainsi que les valeurs extrêmes par des points.

Figure 7. Un exemple de diagramme en boîte.



Les caractéristiques se référant à des tendances centrales

L'écart absolu moyen

L'écart absolu moyen (noté e_M) est la moyenne arithmétique des écarts absolus par rapport à la médiane ou à la moyenne.

Avec x_i valeur des observations, $\bar{x}$ moyenne des observations, M_e la médiane, e_M écart absolu moyen, l'écart absolu moyen des écarts par rapport à la médiane se calcule selon la formule suivante :

$$e_M(M_e) = \frac{1}{n}\sum_{i=1}^{k} n_i |x_i - M_e| = \sum_{i=1}^{k} f_i |x_i - M_e|$$

Et l'écart absolu moyen à la moyenne :

$$e_M(\bar{x}) = \frac{1}{n}\sum_{i=1}^{k} n_i |x_i - \bar{x}| = \sum_{i=1}^{k} f_i |x_i - \bar{x}|$$

Tableau 35. Répartition des SAU viticoles : écart absolu moyen à la médiane (24,1).

Classes (en ha)	c_i	f_i	$\lvert c_i - M_e \rvert$	$f_i \lvert c_i - M_e \rvert$
[0 ; 5[	2,5	13,0	21,6	179,9
[5 ; 10[	7,5	16,0	14,1	352,3
[10 ; 20[	15,0	22,0	4,1	75,0
[20 ; 50[	35,0	34,0	13,4	447,0
[50 ; 100[	75,0	12,0	50,9	593,9
[100 ; 170]	135,0	3,0	100,9	336,4
		100,0		1984,5

$$e_M(M_e) = \sum_{i=1}^{k} f_i |x_i - M_e| = \frac{1984{,}5}{100} \cong 19{,}8 \text{ ha}$$

Tableau 36. Répartition des SAU viticoles : écart absolu moyen à la moyenne (31,8).

Classes (en ha)	c_i	f_i	$\lvert c_i - \bar{x} \rvert$	$f_i \lvert c_i - \bar{x} \rvert$
[0 ; 5[	2,5	13,0	29,3	244,1
[5 ; 10[	7,5	16,0	21,8	544,8
[10 ; 20[	15,0	22,0	11,8	216,2
[20 ; 50[	35,0	34,0	5,7	190,3
[50 ; 100[	75,0	12,0	43,2	504,1
[100 ; 170]	135,0	3,0	93,2	310,7
		100,0		2010,1

$$e_M(M_e) = \sum_{i=1}^{k} f_i \left| x_i - \overline{x} \right| = \frac{2010{,}1}{100} \cong 20{,}1 \text{ ha}$$

Nous pouvons noter que les deux écarts absolus moyens sont très proches ce qui n'est pas systématique.

La variance et l'écart type

La variance et l'écart type sont les indicateurs de dispersion les plus utilisés. La variance V(x) est le carré de la moyenne quadratique des écarts à la moyenne arithmétique, quant à l'écart type σ_x, racine de la variance, il est la moyenne quadratique des écarts à la moyenne arithmétique.
La variance s'exprime de la manière suivante :

$$V(x) = \frac{1}{n} \sum_{i=1}^{m} n_i \left(x_i - \overline{x} \right)^2 = \sum_{i=1}^{m} f_i \left(x_i - \overline{x} \right)^2$$

L'écart type ou écart quadratique moyen est la racine de la variance :

$$\sigma_x = V(x)^{\frac{1}{2}} = \sqrt{V(x)}$$

$$\sigma_x = \sqrt{\sum_{i=1}^{k} f_i \left(x_i - \overline{x} \right)^2}$$

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^{k} n_i \left(x_i - \overline{x} \right)^2}$$

Le calcul de la variance à l'aide de la formule de définition est peu commode, l'utilisation des formules suivantes facilite les calculs.

$$V(x) = \frac{1}{n} \sum_{i=1}^{m} n_i x_i^2 - \overline{x}^2 = \sum_{i=1}^{m} f_i x_i^2 - \overline{x}^2$$

L'écart type est : $\sigma_x = \sqrt{V(x)} = \sqrt{\sum_{i=1}^{m} f_i x_i^2 - \overline{x}^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{m} n_i x_i^2 - \overline{x}^2}$

Le coefficient de variation

Il est défini par le rapport de la moyenne arithmétique à l'écart type, c'est un nombre sans dimension. Il varie comme la dispersion autour de la moyenne, plus il est important et moins la moyenne est significative pour décrire la distribution.

$$CV = \frac{\sigma_x}{\overline{x}}$$

– Calcul de la variance pour une variable discrète

Tableau 36. Ménages selon le nombre de personnes (en milliers) 2010.

Nombre de personnes	x_i	n_i	f_i (en %)	$f_i x_i$	$f_i x_i^2$
1	1	9216,2	34,0	34,0	34,0
2	2	8964,2	33,1	66,1	132,3
3	3	3924,2	14,5	43,4	130,3
4	4	3308,4	12,2	48,8	195,3
5	5	1234,8	4,6	22,8	113,9
6 et plus	6,5	458,7	1,7	11,0	71,5
		27106,5	100	226,2	677,2

$$V(x) = \sum_{i=1}^{m} f_i x_i^2 - \bar{x}^2 = \sum_{i=1}^{m} f_i x_i^2 - \left(f_i x\right)_i^2 = \frac{677,2}{100} - \left(2,262\right)^2 \cong 11,89$$

$$\sigma_x = \sqrt{V(x)} = \sqrt{11,88758} \cong 3,45 \text{ personnes}$$

$$CV = \frac{\sigma_x}{\bar{x}} = \frac{3,45}{2,262} \cong 1,5$$

Exemple : Calcul de la variance pour une variable classée

Tableau 37. Répartition des SAU viticoles.

Classes (en ha)	c_i	f_i	$f_i c_i$	$f_i c_i^2$
[0 ; 5[	2,5	8,3	20,8	52,1
[5 ; 15[	10,0	25,0	250,0	2500,0
[15 ; 25[	20,0	18,3	366,7	7333,3
[25 ; 50[	37,5	33,3	1250,0	46875,0
[50 ; 100[	75,0	11,7	875,0	65625,0
[100 ; 150]	125,0	3,3	416,7	52083,3
		100,0	3179,2	174468,8

$$V(c) = \sum_{i=1}^{m} f_i c_i^2 - \bar{c}^2 = \frac{174468,8}{100} - \left(31,792\right)^2 \cong 733,98$$

$$\sigma_c = \sqrt{V(c)} = \sqrt{733,9774306} \cong 27,1 \text{ ha}$$

$$CV = \frac{\sigma_x}{\bar{x}}$$

$$CV = \frac{\sigma_x}{\bar{x}} = \frac{27,09}{31,79} \cong 0,85$$

Les coefficients de dispersions absolus ou relatifs plus particulièrement les seconds prennent toute leur signification dans le cas des comparaisons entre diverses distributions.

– **Calcul de la variance pour une population composée de sous-populations**

La variance d'une population P composée de h sous-populations P_j s'exprime en fonction des variances et des moyennes des sous-populations. La moyenne arithmétique d'un nombre h de sous-populations ayant pour moyenne $\overline{x}_j$ est définie par :

$$\overline{x} = \sum_{j=1}^{h} f_j \overline{x}_j$$

avec $f_j = \frac{n_j}{n}$ l'importance relative de la sous-population dans l'ensemble de la population.

$$V(x) = \sum_{j=1}^{h} f_j V_j(x) + \left[\sum_{j=1}^{h} f_j \overline{x}_j^2 - \overline{x}^2 \right] = \frac{1}{n} \sum_{j=1}^{h} n_j V_j(x) + \left[\frac{1}{n} \sum_{j=1}^{h} n_j \overline{x}_j^2 - \overline{x}^2 \right]$$

La variance totale est la somme de la moyenne arithmétique des variances et de la variance des moyennes arithmétiques.

La moyenne des variances $\frac{1}{n} \sum_{j=1}^{h} n_j V_j(x)$ est appelée la variance intragroupe.

La variance des moyennes, $\frac{1}{n} \sum_{j=1}^{h} n_j \overline{x}_j^2 - \overline{x}^2$, est appelée la variance intergroupe.

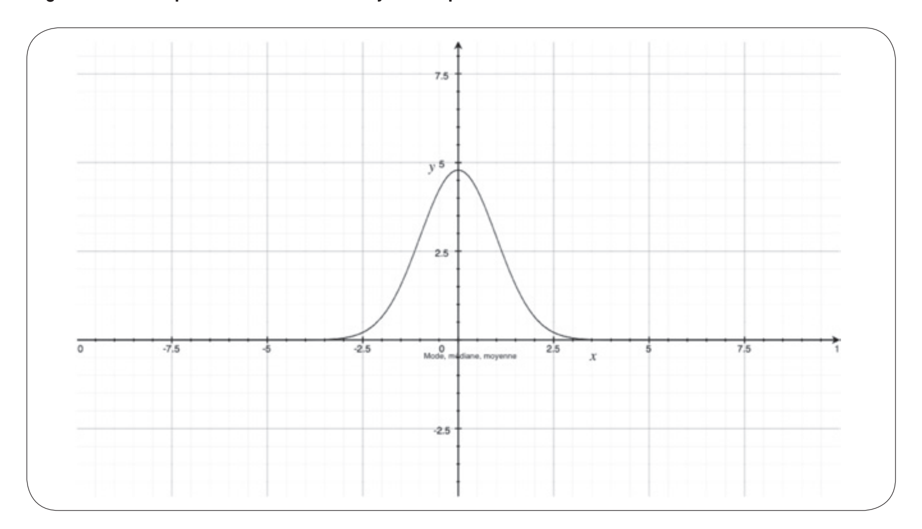
Cette analyse élémentaire de la variance est une première étape dont les extensions au sein du vaste champ de l'analyse des données forment la base pour des analyses statistiques plus développées d'usage fréquent. Ces grandeurs seront très utilisées dans les calculs de la statistique inférentielle qui ne font pas l'objet de cet ouvrage.

La dissymétrie

Deux séries statistiques peuvent avoir la même moyenne et le même écart type sans pour cela être identiques. L'un comme l'autre ne rendent pas compte de la dissymétrie de la distribution. Une estimation de celle-ci est parfois utile. Dans une première étape, il s'agira de reconnaître la dissymétrie, avant de lui donner une valeur précise. Elle sera alors comparée avec la loi normale.

La dissymétrie est évaluée par divers indicateurs. Les coefficients d'asymétrie mesurent la répartition des valeurs de part et d'autre d'une valeur centrale.

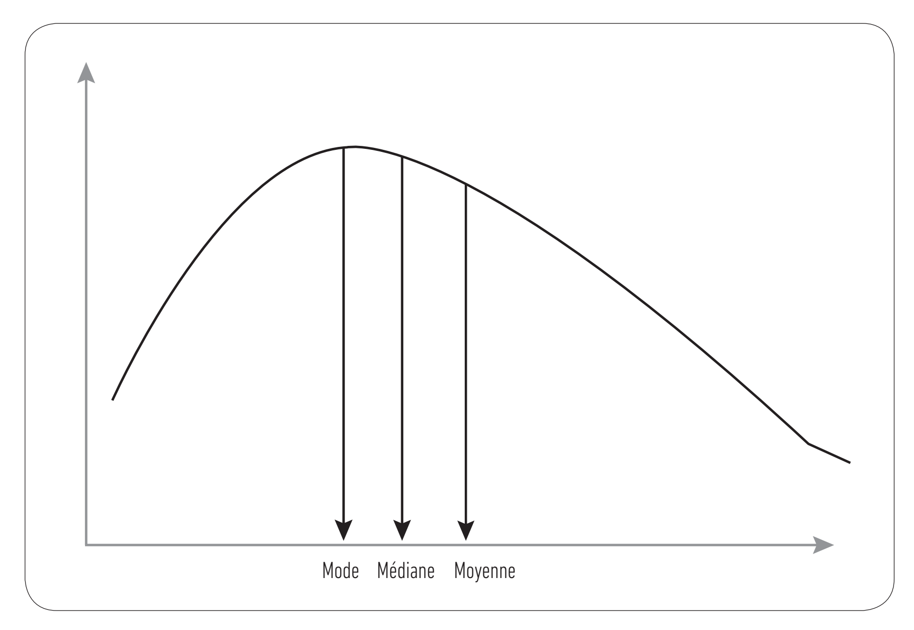

Figure 8. Exemple de distribution symétrique : courbe de la loi normale.



En cas de parfaite symétrie, le mode, la médiane et la moyenne sont identiques.
$M_o = M_e = \overline{x}$
Une première façon simple de considérer la dissymétrie est de comparer les valeurs des tendances centrales. En cas de parfaite symétrie, le mode, la médiane et la moyenne sont identiques. Si le mode est inférieur à la moyenne, la distribution sera oblique à gauche ou étalée à droite.
$M_o < M_e < \overline{x}$
Pour une courbe oblique à droite, un étalement vers la gauche, la relation devient : $M_o > M_e > \overline{x}$

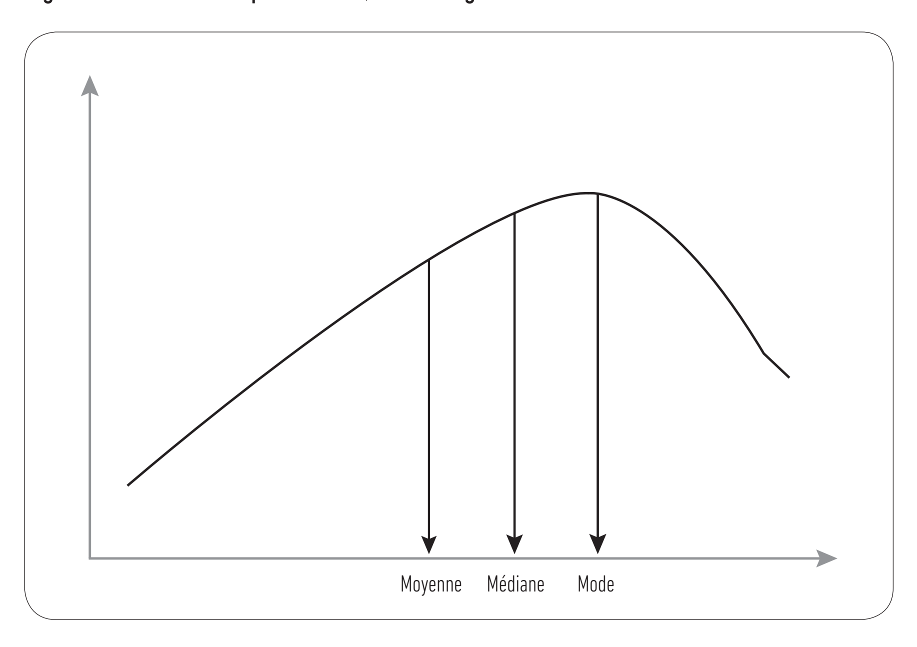
Cela correspond à l'exemple de la distribution des SAU viticoles, dont la moyenne (31,8 ha) est supérieure à la médiane (24,1 ha) qui elle-même est supérieure au mode (10,6 ha) ; la distribution est étalée vers la droite comme nous l'indiquait l'histogramme. La courbe ci-dessous sera dite étalée à droite ou asymétrique à gauche.

Figure 9. Courbe oblique à gauche, étalée à droite.



Pour une courbe oblique à droite, un étalement vers la gauche, nous aurons la relation suivante : $M_o > M_e > \overline{x}$

Figure 10. Courbe oblique à droite, étalée à gauche.



La relation pour une courbe étalée à gauche est : $\overline{x} < M_e < M_o$

Les coefficients de dissymétrie

Plusieurs coefficients permettent d'estimer la dissymétrie d'une distribution. Les coefficients d'asymétrie les plus simples mesurent la répartition des valeurs de part et d'autre d'une valeur centrale, d'autres nécessitent des calculs supplémentaires. Les coefficients les plus utilisés en statistique descriptive sont les moins complexes.
Les calculs des coefficients utiliseront l'exemple de la distribution des SAU viticoles :

Tableau 38. Répartition des SAU viticoles.

Classes (en ha)	c_i	f_i	$f_i c_i$	$f_i c_i^2$	$f_i\left(c_i-\overline{c}\right)^3$
[0 ; 5[	2,5	8,3	32,5	52,1	-209435,8404
[5 ; 15[	10,0	25,0	120,0	2500,0	-258708,8885
[15 ; 25[	20,0	18,3	330,0	7333,3	-30058,48006
[25 ; 50[	37,5	33,3	1190,0	46875,0	6200,214603
[50 ; 100[	75,0	11,7	900,0	65625,0	941129,385
[100 ; 150]	125,0	3,3	405,0	52083,3	2699249,145
		100,0	2977,5	174468,8	3148375,5

Un premier coefficient mesure l'écart relatif du mode et de la moyenne à un indicateur de dispersion. Si nous retenons l'écart type, nous obtenons le premier *coefficient de dissymétrie de Pearson* : D_1.

$$D_1 = \frac{\text{moyenne - mode}}{\text{écart type}} = \frac{\overline{x} - M_0}{\sigma_x}$$

D_1 est un nombre sans dimension.

- $D_1 = 0$ la courbe est symétrique par rapport au mode
- $D_1 > 0$ la courbe est étalée à droite
- $D_1 < 0$ la courbe est étalée à gauche.

Dans l'exemple de la distribution des SAU viticoles :

$$D_1 = \frac{\overline{x} - M_0}{\sigma_x} = \frac{31{,}8 - 10{,}6}{27{,}1} \cong 0{,}78$$

La distribution est étalée à droite.

Le *second coefficient de Pearson* (D_2) estime l'asymétrie par le rapport de l'écart entre la moyenne et la médiane à l'écart type.

$$D_2 = \frac{3(moyenne - médiane)}{écart\ type} = 3\frac{\overline{x} - M_e}{\sigma_x}$$

Pour une distribution symétrique, D_2 est nul, pour une distribution étalée vers la droite D_2 est positif, dans le cas inverse D_2 est négatif.
Pour la distribution des SAU viticoles nous lirons :

$$D_2 = 3\frac{\overline{c} - M_e}{\sigma_c} = 3\frac{31{,}8 - 24{,}1}{27{,}1} = 0{,}85\ .$$

Le *coefficient d'asymétrie de Yule et Kendall*, que l'on nomme s, ne nécessite que la connaissance des trois quartiles :

$$s = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)}\ .$$

Avec :

- s = 0 la distribution est symétrique
- s > 0 la distribution est oblique à gauche (étalée vers la droite)
- s < 0 la distribution est oblique à droite (étalée vers la gauche)

Dans l'exemple de la distribution des SAU viticoles :

$$s = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} = \frac{(42{,}5 - 24{,}1) - (24{,}1 - 10{,}6)}{(42{,}5 - 24{,}1) + (24{,}1 - 10{,}6)} = \frac{4{,}9}{31{,}9} \cong 0{,}15\ .$$

Ce coefficient indique également une distribution étalée vers la droite.
Le *coefficient de Fischer* :

Il est plus algébrique dans sa conception que les précédents, il fait intervenir des écarts à la puissance 3. L'usage de cet indicateur est réservé à des modèles économiques plus complexes que les précédents.

$$\gamma_1 = \frac{\sum_{i=1}^{k} f_i \left(x_i - \overline{x}\right)^3}{\left[\sum_{i=1}^{k} f_i \left(x_i - \overline{x}\right)^2\right]^{3/2}}\ .$$

Avec :

- $\gamma_1 = 0$ distribution symétrique
- $\gamma_1 > 0$ distribution étalée à droite
- $\gamma_1 < 0$ distribution étalée à gauche

Le dénominateur de cette formule est le cube de l'écart type.

$$\gamma_1 = \frac{\sum_{i=1}^{k} f_i \left(c_i - \overline{c}\right)^3}{\left[\sum_{i=1}^{k} f_i \left(c_i - \overline{c}\right)^2\right]^{3/2}} = \frac{3148{,}755}{27{,}1^3} \cong 0{,}16$$

Cet indicateur confirme les conclusions des indicateurs précédents.

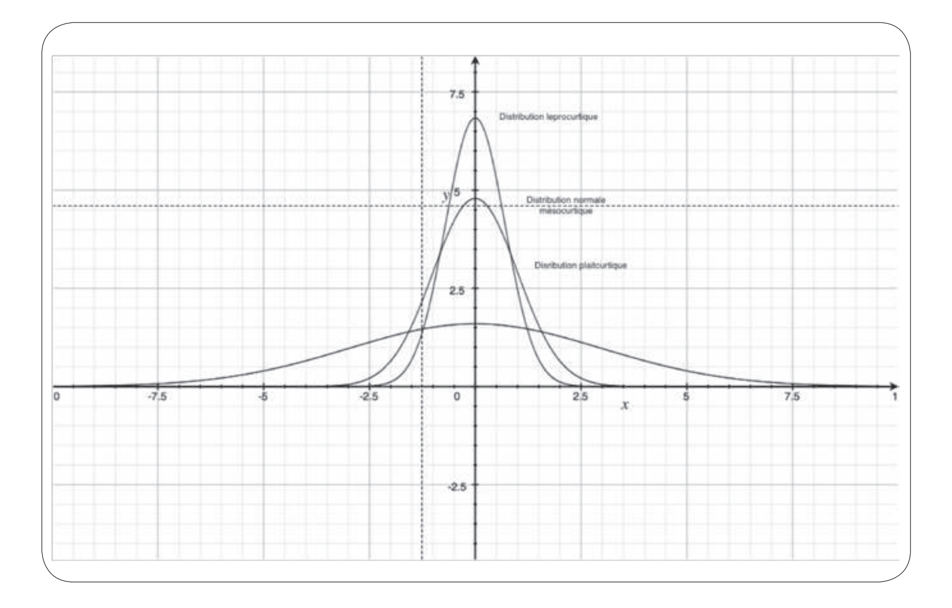
L'aplatissement

Les coefficients d'aplatissement comparent l'allure de la distribution à celle d'une distribution de Laplace-Gauss dit aussi normale. La distribution sera dite leptokurtique si elle est plus « pointue », et platikurtique dans le cas contraire. Des coefficients permettent d'estimer la plus ou moins grande différence avec la loi normale. Une approche graphique est le plus souvent largement suffisante pour les besoins d'une analyse descriptive. Le coefficient d'aplatissement de Yule permet de quantifier cet aplatissement :

$$\gamma_2 = \frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^4}{\sigma^4} - 3 \cdot$$

Si le coefficient est nul la courbe est normale. S'il est négatif, la courbe est platikurtique, s'il est positif la courbe est leptokurtique.

Figure 11. Une illustration des aplatissements.



Les mesures de la concentration

Les mesures de la concentration sont multiples, les plus utilisées sont l'indice C_x, l'indice d'Herfindahl et l'indice de Gini associé à la courbe de Lorenz.

C_x

Cet indicateur, très utilisé, est facile à calculer et d'interprétation aisée. Il mesure les fréquences cumulées des valeurs de la variable pour les x premières entreprises du domaine considéré. Le C_x représente l'importance relative des x premières entreprises. Si les quatre plus grandes entreprises du secteur de l'énergie font ensemble 75 % du chiffre d'affaires des entreprises du secteur, le C_4 du secteur énergie, pour les chiffres d'affaires, sera de 75.

Si, m¡ représente la part de marché, ou du chiffre d'affaires HT, ou de l'effectif salarié de l'entreprise i alors $C_x = \sum_{i=1}^{x} m_i$.

Le tableau ci-dessous illustre l'utilisation courante de cet indicateur.

Tableau 39. Classements des 20 premiers constructeurs automobiles dans le monde en 2012 en milliers de voitures vendues.

Classement	Nom	Part du marché mondial (%)	C_x
1	Toyota	10,9	10,9
2	GM	10,7	21,6
3	Volkswagen	9,3	30,9
4	Hyundai	7,3	38,2
5	Ford	6,3	44,5
6	Nissan	5,0	49,6
7	Honda	4,6	54,2
8	PSA	4,6	58,8
9	Suzuki	3,7	62,5
10	Renault	4,7	67,2
11	Fiat	3,1	70,2
12	Daimler	2,5	72,7
13	Chrysler	2,0	74,7
14	BMW	1,9	76,6
15	Mazda	1,7	78,2
16	Mitsubishi	1,5	79,7
17	Chana Automobile	1,4	81,1
18	Tata	1,3	82,4
19	Groupe FAW	1,1	83,5
20	Zheijiang Gely	1,0	84,5

http://www.actualitix.com/

Le $C_1 = 10{,}9$ signifie que le premier groupe automobile représente 10,9 % de la production automobile mondiale ; $C_5 = 44{,}5$ indique que les cinq premiers fournissent 44,5 % et les vingt premiers constructeurs assurent 84,5 de la production mondiale $C_{20} = 84{,}5$. Ce secteur est très concentré au plan mondial. La mesure fournie ici sous-estime la concentration puisque Renault et Nizan sont en partenariat ainsi que Fiat et Chrysler, sans évoquer des accords spécifiques (production en commun de moteurs) entre certains constructeurs.
L'avantage de cet indicateur est de nécessiter peu d'informations statistiques. Il suffit de connaître la production totale d'un secteur et la production totale ou le chiffre d'affaires total des x premières unités. Par contre, il ne tient pas compte de n -x autres unités. Cet indice est utilisé pour construire une typologie des marchés.

Une typologie des marchés

Les parts de marché sont le plus souvent estimées par les CA des entreprises. La classification suivante est souvent utilisée afin de caractériser certaines situations[3] :

- monopole, la première firme occupe plus de 80 % du marché $C_1 \geq 80$;
- firme leader, la première firme réalise entre 50 % et 80 % du marché, les autres firmes sont de taille négligeable $50 \leq C_1 \leq 80$;
- duopole, deux firmes se partagent à peu près également 80 % du marché $C_2 \cong 80$;
- oligopole asymétrique, trois ou quatre firmes réalisent 80 % du marché, dont environ 40 % pour la première ;
- oligopole symétrique est constitué de trois ou quatre unités se partageant de façon équivalente 80 % du marché $C_4 \cong 80$ et $C_1 = 40$;
- concurrence asymétrique, la première firme réalise entre 20 % et 50 % du marché, les autres ont des tailles très inférieures $20 \leq C_1 \leq 40$;
- concurrence totale, la première firme occupe moins de 20 % du marché $C_1 \leq 20$.

Il existe des formes de concentration du côté des acheteurs

- le monopsone : un petit nombre de producteurs, un seul acheteur. Le marché du tabac brut est donc un monopsone.
- l'oligopsone : un très grand nombre de producteurs, un petit nombre d'acheteurs. Les centrales d'achat des magasins à grande surface forment un oligopsone.

3. *Les groupes de sociétés dans le système productif français*, Paris, INSEE.

L'indice Hirschman[4]-Herfindahl (indice HH)

L'indice C_x ne permet pas toujours de différencier, sauf de manière qualitative, des situations pourtant fort différentes. Par exemple le C_4 des quatre distributions suivantes est de 90 %.

Tableau 40. Distributions des parts de marché pour les quatre premières entreprises.

D I (en % de part de marché)	D II (en % de part de marché)	D III (en % de part de marché)	D IV (en % de part de marché)
77	42	40	25
5	38	20	25
4	6	16	20
4	4	14	20
90	90	90	90

À la lecture de ce tableau, il apparaît que les quatre distributions ne sont pas équivalentes en termes de répartition des pouvoirs de marché. La distribution D-I correspond à une situation de firme leader, la situation D-II à un duopole, la situation D-III à un oligopole asymétrique, la distribution D-IV à oligopole symétrique.

L'indice Herfindahl-Hirschman (HH) permet de mettre en évidence ces différentes situations. Il est calculé comme la somme des carrés des parts de marché de toutes les entreprises :

$$C_H = \sum_{i=1}^{n} \left[\frac{X_i}{X} \right]^2 = \sum_{i=1}^{n} m_i^2 \cdot$$

Avec X_i part de la variable que représente l'entreprise i, $X = \sum_{i=1}^{n} X_i$ et où m_i est la part de marché détenue par l'entreprise i.

L'indice HH est compris entre 0 et 10 000 si l'importance de chaque entreprise est exprimée en pourcentage. Il croît avec la concentration d'où son utilisation fréquente pour mesurer le niveau de concurrence existant au sein d'une zone géographique.

4. Albert Hirschman économiste américain d'origine allemande (1915-2012), Orris Herfindahl (1918-1972).

Tableau 41. Indices de concentration de Herfindahl.

Rangs des entreprises	D I (en % de part de marché)	D II (en % de part de marché)	D III (en % de part de marché)	D IV (en % de part de marché)
1	77	42	40	25
2	5	38	20	25
3	4	6	16	20
4	4	4	14	20
$\sum_{i=1}^{4} m_i^2$	5 986	3 260	2 452	2 050

Le marché correspondant à la distribution I est deux fois plus concentré que celui de la distribution IV, alors que les concentrations semblaient identiques en utilisant l'indicateur C_x.

Pour décrire un marché à un instant donné, il est souvent plus éloquent de se référer au nombre équivalent d'Herfindahl η_H . Il représente le nombre d'entreprises, de taille identique, qui réaliseraient la même valeur de concentration que celle donnée par le C_H.

$$\eta_H = \frac{1}{C_H}$$

Tableau 42. Nombre équivalent de Herfindahl.

	D I	D II	D III	D IV
$\sum_{i=1}^{4} m_i^2$	5 986	3 260	2 452	2 050
η_H	1,7	3,1	4,1	4,9

Le nombre équivalent montre bien la diversité des structures de marché de ces quatre distributions. L'information est identique à celle fournie par l'indice HH, elle est plus immédiate et donc plus éloquente.
Le Département de la justice des États-Unis considère qu'un indice HH inférieur à 1000 (1500 pour les agences) indique un marché de concurrence. Pour un indice HH compris entre 1000 et 1800 (1500, 2500 pour les agences), le marché est modérément concentré au-delà de 1800 (2500 pour les agences) le marché est hautement concentré. Une fusion qui augmente l'indice HH de plus de 100 est passible de la loi *antitrust* pour un secteur hautement concentré, une fusion augmentant l'indice HH de plus 200 points nécessite une intervention.

L'indice HH est calculé pour mesurer la concentration de divers domaines, celui du mixte énergétique est une possibilité de comparaison de situation nationale.

$$\text{HHmé}=\sum_{j=1}^{n} S_j^2$$

avec j le type d'énergie et S la part de l'énergie j dans le total de la consommation énergétique.

Tableau 43. Structure de la consommation d'énergie par type.

2010	**France**		**Allemagne**	
	S_j	S_j^2	S_j	S_j^2
Gaz	16	256	22	484
Pétrole	31	961	34	1156
Nucléaire	41	1681	11	121
Renouvelables	8	64	10	100
Combustibles solides	4	16	23	529
Total	100	2978	100	2390

Occasional Papers 145 | April 2013 Member States' Energy Dependence: An Indicator-Based Assessment <ec.europa.eu/economy_finance/publications>

Le niveau de concentration de la France est de 0,30 (0,2978) et il est supérieur à celui de l'Allemagne qui affiche 0.24 (0,2390) en 2010. Le niveau de l'UE à 27 est de 0,24. Le mixte énergétique de la France est moins diversifié que celui de l'Union européenne et de l'Allemagne, ce résultat provient vraisemblablement de l'importance de l'énergie nucléaire en France.

Les PIB régionaux sont connus, l'indice HH permet de donner une évaluation de l'évolution de la concentration des PIB entre les régions et de calculer le nombre de régions ayant la même importance pour cet indicateur.

Tableau 44. Produits intérieurs bruts régionaux (PIBR) en valeur en millions d'euros.

	Part des PIB régionaux			
	m_i (%)		m_i^2	
	1990	2011	1990	2011
Alsace	2,9	2,7	8,6	7,3
Aquitaine	4,3	4,5	18,6	20,4
Auvergne	2,0	1,7	3,9	2,9
Bourgogne	2,5	2,2	6,4	4,6
Bretagne	3,9	4,2	15,6	17,4
Centre	3,9	3,4	15,1	11,4
Champagne-Ardenne	2,2	1,9	4,8	3,5
Corse	0,3	0,4	0,1	0,2
Franche-Comté	1,7	1,5	2,8	2,1
Ile-de-France	28,6	30,6	817,6	935,8
Languedoc-Roussillon	2,9	3,2	8,7	10,3
Limousin	1,0	0,9	1,1	0,8
Lorraine	3,4	2,9	11,6	8,2
Midi-Pyrénées	3,7	4,0	14,0	15,9
Nord-Pas-de-Calais	5,4	5,2	29,7	26,9
Basse-Normandie	2,0	1,8	4,1	3,3
Haute-Normandie	2,8	2,5	8,0	6,4
Pays de la Loire	4,6	5,1	20,8	25,5
Picardie	2,7	2,3	7,4	5,2
Poitou-Charentes	2,3	2,3	5,2	5,1
Provence-Alpes-Côte d'Azur	7,0	7,1	49,0	51,1
Rhône-Alpes	9,6	9,9	93,0	97,1
France métropolitaine	100,0	100	1145,8	1261,4

Source : Insee, base 2005

L'indice HH ($C_H = \sum_{i=1}^{n} m_i^2$) augmente entre 1900 et 2011 de plus de 100 points indiquant une plus grande inégalité entre les régions.

En calculant le nombre équivalent de région ($\eta_H = \frac{1}{C_H}$) et en arrondissant à l'unité ; le nombre de régions ayant un même PIB passe de 9 en 1990 à 8 en 2011 traduisant une concentration significative de la richesse nationale pour quelques régions.

L'indice de Gini et la courbe de Lorenz

L'indice de Gini associé à la courbe de Lorenz mesure l'importance de la concentration. La courbe de concentration de Lorenz visualise graphiquement la concentration. La construction de cette courbe nécessite la connaissance de la distribution des effectifs et de celle de la variable.

La courbe de Lorenz

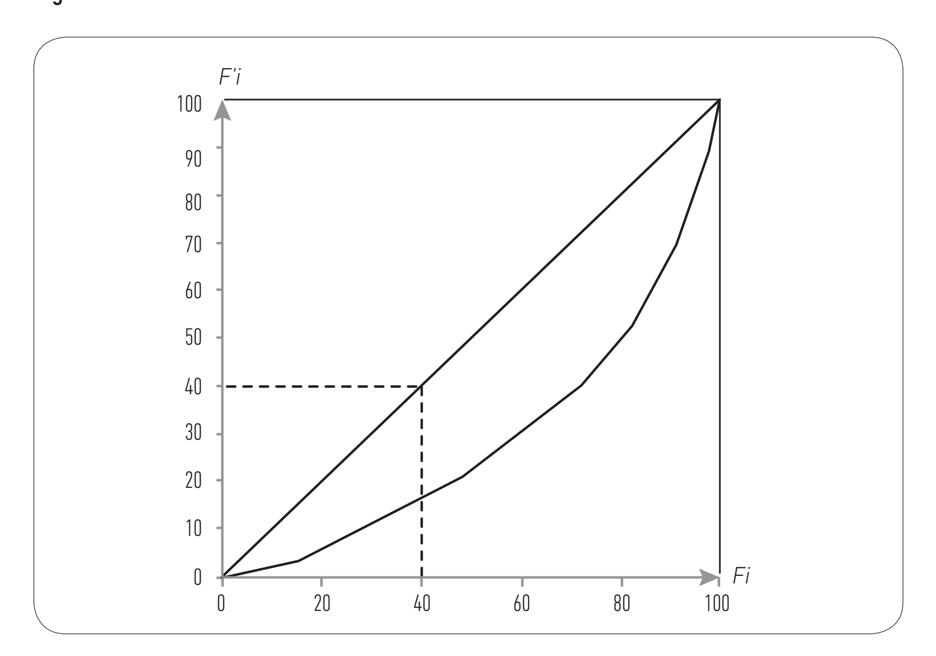
La courbe se construit en inscrivant en abscisses le pourcentage F_i cumulé des effectifs et en ordonnée le pourcentage cumulé F'_i des valeurs la variable statistique.
Les calculs nécessaires pour obtenir les différentes caractéristiques sont grandement facilités par la construction du tableau statistique dont un modèle est donné ci-dessous.

Tableau 45. Courbe de Lorenz et indice de Gini.

Valeurs de la variable	**Fréquences**	**Fréquences cumulées**				
x_i ou c_i (1)	f_i	F_i	$f_i x_i$	f'_i	F'_i	$f_i(F'_{i-1}+F'_i)$
x_1	f_1	$F_1=f_1$	$f_1 x_1$	$f'_1 = \frac{f_1 x_1}{\sum_{i=1}^{m} f_i x_i}$	$F'_1 = f'_1$	$f_1 F'_1$
x_i	$f_i = \frac{n_i}{n}$	$F_i = \sum_{k=1}^{i} f_k$	$f_i x_i$	$f'_i = \frac{f_i x_i}{\sum_{i=1}^{m} f_i x_i}$	$F'_k = \sum_{k=1}^{i} f'_k$	$f_i(F'_{i-1}+F'_i)$
x_m	f_m	$F_m=1$	$f_m x_m$	$f'_m = \frac{f_m x_m}{\sum_{i=1}^{m} f_i x_i}$	$F'_{\substack{1\\m}} = 1$	$f_m(F'_{m-1}+F'_m)$
Total	$\sum_{i=1}^{m} f_i = 1$		$\sum_{i=1}^{m} f_i x_i$	$\sum_{i=1}^{m} f'_i = 1$		$\sum_{i=1}^{m} f_i\left(F'_{i-1}+F'_i\right)$

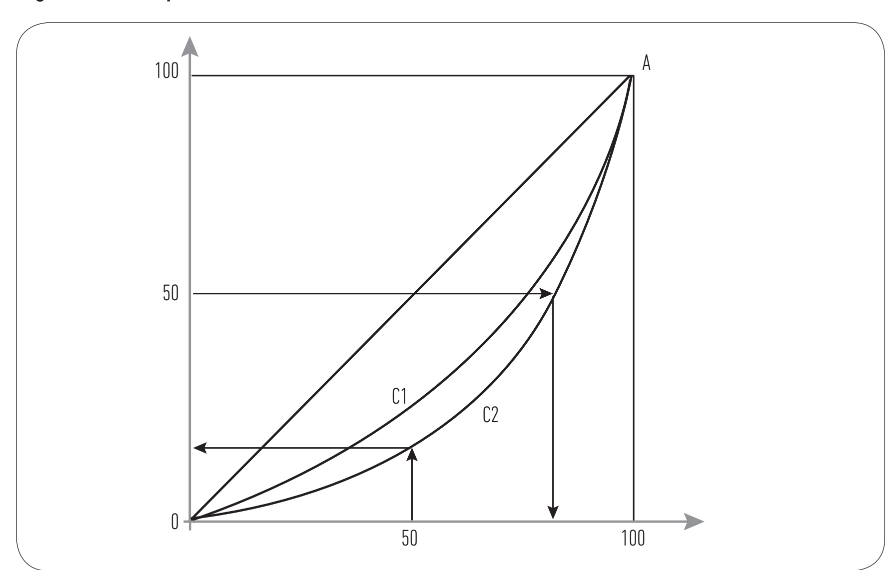
(1) selon que la variable est ou non classée

Figure 12. Courbe de Lorenz.



La droite OA indique l'équipartition. En effet, par exemple, à 40 % de l'effectif correspond 40 % de la valeur totale de la variable. Plus la courbe est proche de OA et plus le niveau de concentration sera faible. La courbe de concentration permet de déterminer sur un même graphique la position relative de la médiane et de la médiale. La comparaison entre ces deux tendances centrales permet une première mesure de la concentration. Une première mesure s'appuie sur l'écart entre la médiale et la médiane plus il est important plus la concentration est forte. Cet écart est mesuré par $\Delta M = M_l - M_e$. Une seconde méthode est de comparer l'écart à l'étendue de la distribution ou un écart interquantile comme l'écart interdécile.

Figure 13. Comparaison de deux courbes de Lorenz.



La courbe C_1 représente une distribution ayant une concentration plus faible que la distribution représentée par la courbe C_2. Il est à remarquer, sur le graphique, que plus la variable est concentrée et plus la médiale est supérieure à la médiane. En revanche, lorsque la concentration est nulle, la médiane et la médiale sont confondues.

Plus la courbe de Lorenz est proche de la diagonale moins l'inégalité des revenus est importante.
Cette détermination graphique est très simple ; il sera ensuite nécessaire de rechercher sur un tableau des fréquences la classe correspondant au pourcentage des effectifs ou des valeurs cumulées.
La représentation de distributions par des courbes de Lorenz ne permet pas toujours de comparer la concentration de deux distributions.

Les aires OC1A et OC2A semblent égales ; néanmoins, les deux distributions sont différentes. C'est pourquoi une mesure utilisant l'indice de Gini permet tout à la fois de faciliter les comparaisons et de quantifier l'inégalité.

L'indice de Gini

L'indice de Gini associé à la courbe de Lorenz mesure l'importance de la concentration. L'indice de concentration I_G de Gini représente le double de l'aire comprise entre la courbe et la diagonale. L'indice I_G résume la concentration à l'aide d'un seul nombre.

$$I_G = 1 - \sum_{i=1}^{k} f_i \left(F'_{i-1} + F'_i \right)$$

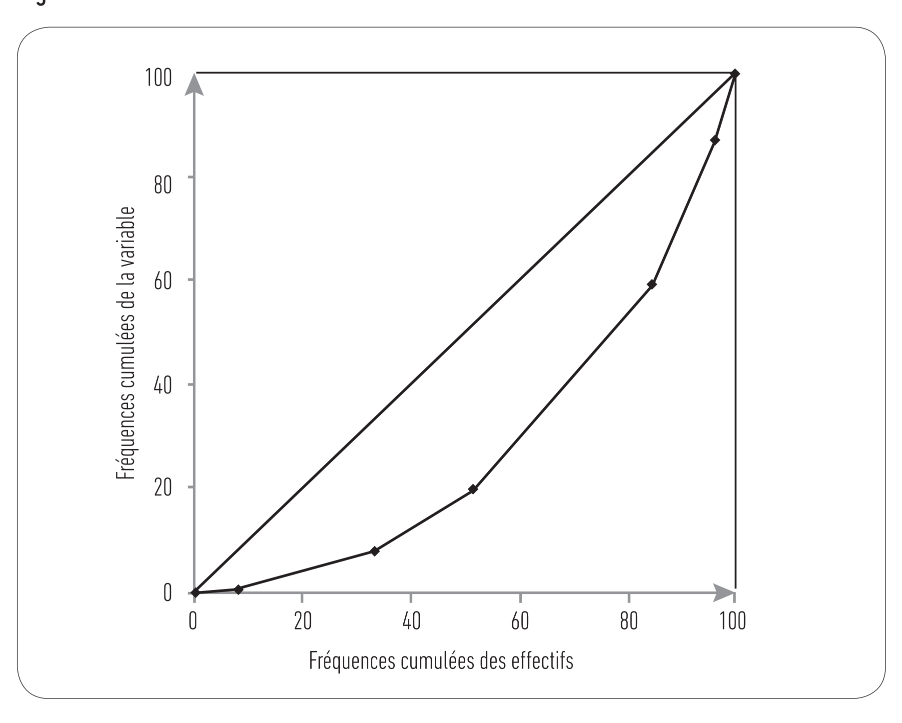
L'indice I_G varie de 0 à 1, plus il est proche de 0 plus la concentration est faible, s'il est proche de 1 la concentration est forte. L'indice de concentration est un nombre sans dimension, indépendant de l'unité choisie.

Exemple avec la distribution des SAU

Tableau 46. Répartition des SAU viticoles.

Classes (en ha)	c_i	f_i	F_i	$f_i c_i$	f'_i	F'_i	$f_i(F'_{i-1}+F'_i)$
[0 ; 5[	2,5	8,3	8,3	20,8	0,7	0,7	5,5
[5 ; 10[	7,5	25,0	33,3	250,0	7,9	8,5	229,4
[10 ; 20[	15,0	18,3	51,7	366,7	11,5	20,1	523,8
[20 ; 50[	35,0	33,3	85,0	1250,0	39,3	59,4	2647,4
[50 ; 100[	75,0	11,7	96,7	875,0	27,5	86,9	1706,4
[100 ; 150]	135,0	3,3	100,0	416,7	13,1	100,0	623,0
		100,0		3179,2	100,0		5735,5

Figure 14. Courbe de Lorenz.



La comparaison entre la médiale et la médiale fournit une première indication de la concentration.

$$M_e = b_i + a_i \cdot \frac{50 - F_{i-1}}{f_i} = 15 + 10 * \frac{50 - 33,3}{18,3} \cong 24,1$$

$$M_l = b_i + a_i \cdot \frac{50 - F'_{i-1}}{f'_i} = 25 + 25 * \frac{50 - 20,1}{39,3} \cong 44,0$$

$$\frac{\Delta M}{étendue} = \frac{19,9}{150} \cong 0,133 \quad \frac{\Delta M}{D_9 - D_1} = \frac{19,9}{68,2 - 6} \cong 0,32$$

Pour calculer l'indice de Gini, il ne faut pas oublier que si les fréquences sont données en pourcentage, il faut diviser le total des $f_i\left(F'_{i-1}+F'_i\right)$ par 100^2.

$$I_G = 1 - \sum_{i=1}^{k} f_i\left(F'_{i-1} + F'_i\right) = 1 - \frac{5735,5}{10000} \cong 0,43$$

Ce chapitre fournit les principaux indicateurs d'usage courant pour synthétiser une distribution statistique. Il est important de garder à l'esprit que tous les résultats dépendent de la qualité des informations et des traitements numériques. La généralisation de l'accès aux appareils de traitement numérique des données autorise la multiplication des calculs et assure la justesse des résultats. Cette nouvelle situation permet de pouvoir se concentrer sur l'analyse des données et des résultats que sur les méthodes de calcul. C'est pourquoi ce chapitre insiste sur les conditions de validité des indicateurs.

Chapitre 3

Les distributions statistiques à deux dimensions

Il est souvent pertinent d'étudier une population à l'aide de plusieurs caractères. Ce chapitre présente les distributions pour lesquelles nous disposons d'observations concernant simultanément deux caractères – qui peuvent être qualitatifs ou quantitatifs – pour chaque individu de la population. Le regroupement des observations dans un tableau d'effectifs ou tableau de contingence facilite la lecture en regroupant les observations par modalités. La mise en évidence d'une relation entre deux variables signifie que leurs évolutions statistiques sont liées.

Le chapitre présente tout d'abord les tableaux de contingence les plus utilisés en économie et les techniques statistiques de production des caractéristiques synthétiques. Il examinera ensuite la covariance et l'estimation et la validité des paramètres liant deux variables dans le cadre d'un ajustement linéaire.

Les tableaux de contingence

Les tableaux de contingence associent deux caractères ou deux variables. Des tableaux plus complexes sont envisageables dont le traitement statistique est plus compliqué et sort du cadre de cet ouvrage.

Les deux tableaux ci-dessous présentent le croisement de variables discrètes et de caractères. Des tableaux croisant tous les types caractères et de variables sont possibles sans modifier les calculs.

Tableau 1. Croisement de deux variables discrètes.

Variable 1 \ Variable 2	y_1	y_j	y_p	Effectif marginal de la variable 1
x_1	n_{11}	n_{1j}	n_{1p}	$n_{1.} = \sum_{j=1}^{p} n_{1j}$
x_i	n_{i1}	n_{ij}	n_{ip}	$n_{i.} = \sum_{j=1}^{p} n_{ij}$
x_m	n_{m1}	n_{mj}	n_{mp}	$n_{m.} = \sum_{j=1}^{p} n_{mj}$
Effectif marginal de la variable 2	$n_{.1} = \sum_{i=1}^{m} n_{i1}$	$n_{.j} = \sum_{i=1}^{m} n_{ij}$	$n_{.p} = \sum_{i=1}^{m} n_{ip}$	$n = \sum_{i=1}^{p} n_{.i} = \sum_{j=1}^{m} n_{j.}$

Tableau 2. Croisement de deux caractères.

Caractère A \ Caractère B	Modalité 1	Modalité j	Modalité p	Effectif marginal du caractère A
Modalité 1	n_{11}	n_{1j}	n_{1p}	$n_{1.} = \sum_{k=1}^{p} n_{1k}$
Modalité i	n_{i1}	n_{ij}	n_{ip}	$n_{i.} = \sum_{k=1}^{p} n_{ik}$
Modalité m	n_{m1}	n_{mj}	n_{mp}	$n_{m.} = \sum_{k=1}^{p} n_{mk}$
Effectif marginal du caractère B	$n_{.1} = \sum_{k=1}^{m} n_{k1}$	$n_{.j} = \sum_{k=1}^{m} n_{kj}$	$n_{.p} = \sum_{k=1}^{m} n_{kp}$	$n = \sum_{k=1}^{p} n_{.k} = \sum_{k=1}^{m} n_{k.}$

L'effectif n_{ij} de la case (i, j) est le sous-ensemble de la population P des individus qui présentent simultanément la modalité A_i et la modalité B_j. La distribution s'écrit $\{(A_i, B_j, n_{ij})\}$. Tous les individus présentant ces deux modalités sont considérés comme équivalents. Le total des lignes et le total des colonnes définissent des distributions marginales. Une ligne ou une colonne constitue une distribution conditionnelle.

Le total des effectifs de la ligne i s'écrit n_i :

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ip}$$

$$n_{i.} = \sum_{j=1}^{p} n_{ij}$$

Le total de la colonne j s'écrit n.j :

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{mj}$$

$$n_{.j} = \sum_{i=1}^{m} n_{ij}$$

Le total n peut se calculer de plusieurs façons :

$$n = \sum_{j=1}^{p} n_{.j} = \sum_{i=1}^{m} n_{i.} = \sum_{i=1}^{m}\sum_{j=1}^{p} n_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{p} n_{ij} .$$

On appelle f_{ij} la fréquence de la modalité (x_i, y_i) ou de l'événement A_i, B_j, la proportion d'individus qui présentent simultanément A_i et B_j soit :

$$f_{ij} = \frac{n_{ij}}{n} .$$

La fréquence f_{ij} est alors dite conjointe.
Pour illustrer ces démonstrations, nous allons étudier le tableau du commerce intra régional et interrégional de marchandises publié par l'OMC en 2012 qui peut être analysé selon plusieurs points de vue.

Tableau 3. Commerce intra régional et interrégional de marchandises, 2012 (tableau des valeurs).

Destination / Origine	Amérique du Nord	Amérique du Sud et Centrale	Europe	CEI	Afrique	Moyen-Orient	Asie	Monde
Amérique du Nord	1 151	217	380	18	38	75	488	2 367
Amérique du Sud et centrale	187	202	128	8	21	17	172	734
Europe	492	124	4 383	245	211	208	643	6 306
Communauté d'États indépendants (CEI)	37	7	430	149	14	20	127	784
Afrique	74	30	240	2	81	17	160	604
Moyen-Orient	118	11	148	7	39	116	732	1 171
Asie	975	196	855	121	177	260	3 012	5 597
Monde	3 035	787	6 564	550	580	714	5 333	17 563

Source : OMC

Le tableau ci-dessus est en valeur, ce qui rend les comparaisons difficiles. Pour en faciliter la lecture, il est possible de produire trois tableaux. Le premier donnera l'importance dans le commerce mondial des échanges

entre les grandes zones (les fréquences f_{ij}). Le second tableau explicitera les fréquences des échanges par zone et par destination (les fréquences $f_{i/j}$) tandis que le troisième livrera la répartition des échanges selon la destination (les fréquences $f_{i/j}$).

Tableau 4. Part des courants d'échanges régionaux dans le commerce mondial de marchandises (en %).

f_{ij} Destination / Origine	Amérique du Nord	Amérique du Sud et Centrale	Europe	CEI	Afrique	Moyen-Orient	Asie	Monde
Amérique du Nord	6,6	1,2	2,2	0,1	0,2	0,4	2,8	13,5
Amérique du Sud et centrale	1,1	1,1	0,7	0,0	0,1	0,1	1,0	4,2
Europe	2,8	0,7	25,0	1,4	1,2	1,2	3,7	35,9
(CEI)	0,2	0,0	2,4	0,8	0,1	0,1	0,7	4,5
Afrique	0,4	0,2	1,4	0,0	0,5	0,1	0,9	3,4
Moyen-Orient	0,7	0,1	0,8	0,0	0,2	0,7	4,2	6,7
Asie	5,6	1,1	4,9	0,7	1,0	1,5	17,2	31,9
Monde	17,3	4,5	37,4	3,1	3,3	4,1	30,4	100,0

Tableau 5. Part des courants d'échanges régionaux dans les exportations totales de marchandises de chaque région (en %).

$f_{j/i}$ Destination / Origine	Amérique du Nord	Amérique du Sud et Centrale	Europe	CEI	Afrique	Moyen-Orient	Asie	Monde
Amérique du Nord	48,6	9,2	16,0	0,8	1,6	3,2	20,6	100,0
Amérique du Sud et centrale	25,4	27,5	17,4	1,1	2,9	2,3	23,4	100,0
Europe	7,8	2,0	69,5	3,9	3,3	3,3	10,2	100,0
CEI	4,7	0,9	54,9	19,0	1,8	2,6	16,2	100,0
Afrique	12,2	5,0	39,8	0,3	13,4	2,8	26,4	100,0
Moyen-Orient	10,1	0,9	12,7	0,6	3,4	9,9	62,5	100,0
Asie	17,4	3,5	15,3	2,2	3,2	4,7	53,8	100,0
Monde	17,3	4,5	37,4	3,1	3,3	4,1	30,4	100,0

Figure 1. Graphique des exportations totales de marchandises de chaque région.

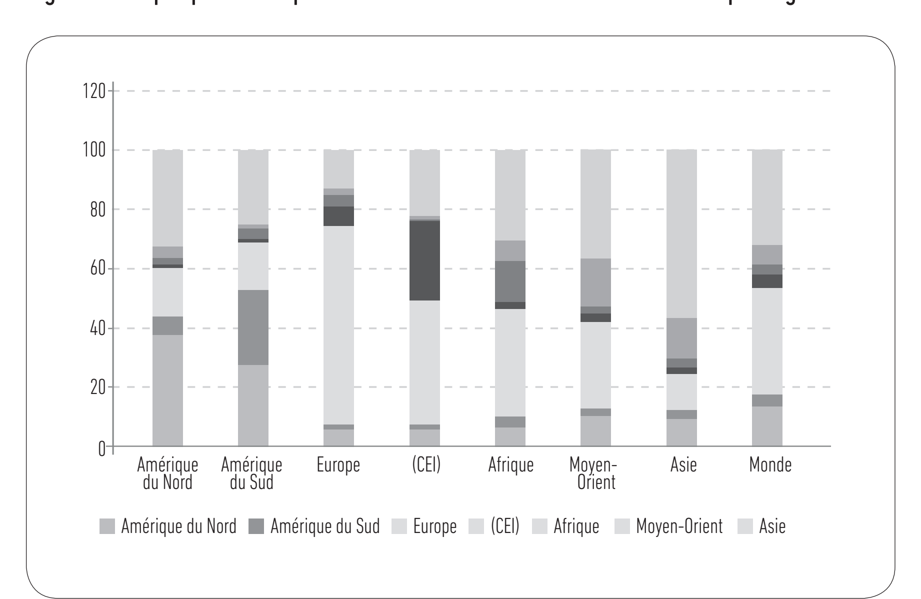
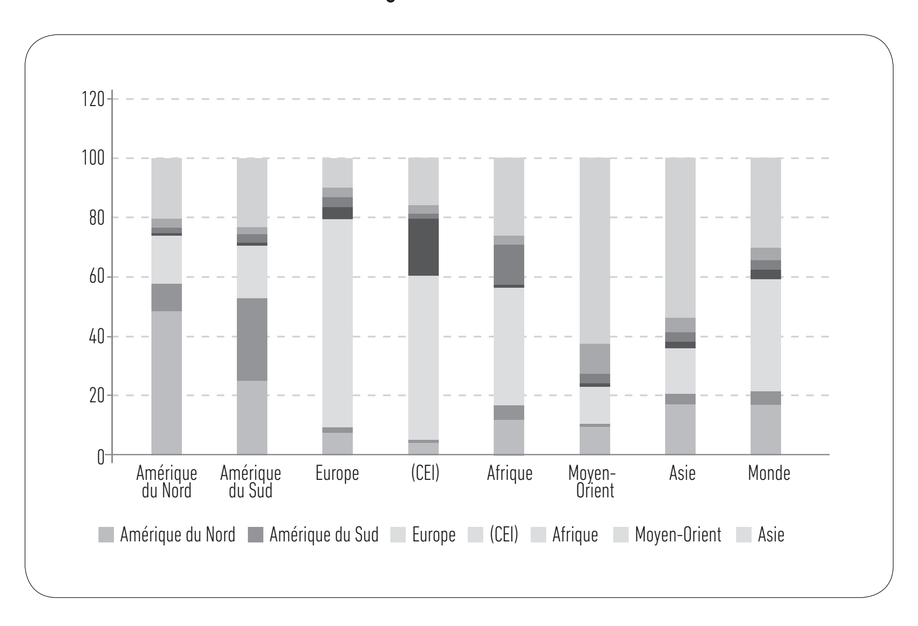


Tableau 6. Part de chaque région dans les exportations mondiales de marchandises à destination de la région.

$f_{j/i}$ Destination / Origine	Amérique du Nord	Amérique du Sud et Centrale	Europe	CEI	Afrique	Moyen-Orient	Asie	Monde
Amérique du Nord	37,9	27,6	5,8	3,3	6,5	10,5	9,2	13,5
Amérique du Sud et centrale	6,2	25,6	1,9	1,5	3,6	2,4	3,2	4,2
Europe	16,2	15,7	66,8	44,6	36,3	29,2	12,0	35,9
(CEI)	1,2	0,9	6,6	27,0	2,4	2,8	2,4	4,5
Afrique	2,4	3,9	3,7	0,3	13,9	2,4	3,0	3,4
Moyen-Orient	3,9	1,4	2,3	1,3	6,8	16,2	13,7	6,7
Asie	32,1	24,9	13,0	21,9	30,4	36,5	56,5	31,9
Monde	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Figure 2. Graphique de la part de chaque région dans les exportations mondiales de marchandises à destination de la région.



Chacun des tableaux et des graphiques donne une image des échanges internationaux irréductibles aux deux autres. Cet exemple illustre les différents points de vue possibles dans l'analyse d'un tableau de contingence et en enrichit la compréhension. Les deux distributions marginales peuvent être individualisées aux marges du tableau et au sein du tableau ce seront les $m+p$ distributions conditionnelles. Pour chacune de ces distributions, il est envisageable de calculer les caractéristiques de synthèse présentées antérieurement. L'accent sera mis sur les calculs de moyennes arithmétiques et de variance qui permettent de mettre en lumière des relations entre les variables.

Les distributions marginales

Les $\{A_i, n_i\}$ et $\{B_j, n_{.j}\}$ constituent les deux distributions marginales du tableau. La distribution marginale du caractère A est définie par les sommes en ligne, c'est une distribution à une seule dimension ; le caractère B n'intervient pas. La fréquence de la modalité A_i est

$$f_{i.} = \sum_{j=1}^{p} f_{ij} = \sum_{j=1}^{p} \frac{n_{ij}}{n} = \frac{n_{i.}}{n}$$

La distribution marginale du caractère B est définie par les sommes en colonne, c'est aussi une distribution à une seule dimension où le caractère A ne joue aucun rôle.

La fréquence de la modalité B_j est

$$f_{.j} = \sum_{j=1}^{m} f_{ij} = \sum_{j=1}^{m} \frac{n_{ij}}{n} = \frac{n_{i.}}{n} \cdot$$

Comme pour les distributions à un caractère, la somme des fréquences est égale à 1.

$$\sum_{i=1}^{m} \sum_{j=1}^{p} f_{ij} = \sum_{j=1}^{p} f_{.j} = \sum_{i=1}^{m} f_{i.} = 1$$

Les distributions conditionnelles

Une ligne ou une colonne constitue une distribution conditionnelle. Soit $n_{i.}$ individus qui possèdent la modalité A_i parmi ceux-ci $\frac{n_{ij}}{n_{i.}}$ présentent également la modalité B_j.

On dit que la fréquence conditionnelle de la modalité B_j liée par A_i est :

$$f_{j/i} = f(B_j / A_i) = \frac{n_{ij}}{n_{i.}}.$$

Les différentes fréquences conditionnelles pour une même modalité A_i définissent la distribution conditionnelle de B liée par A_i $\{B_j, f_{j/i}\}$. C'est une distribution à seul caractère et il y a autant de distributions conditionnelles de B qu'il existe de modalités A_i.

La distribution conditionnelle de A liée par B_j est définie de façon analogue. La fréquence conditionnelle de A_i liée par B_j.

$$f_{i/j} = f(A_i / B_j) = \frac{n_{ij}}{n_{.j}}$$

La distribution conditionnelle de A_i liée par B_j se compose des différentes fréquences conditionnelles, $\{A_i, f_{i/j}\}$. Il existe autant de distributions conditionnelles de A qu'il existe de modalités de B_j.

La relation entre fréquences marginales et conditionnelles

Les fréquences marginales et conditionnelles sont liées par les relations suivantes :

$$f_{ij} = \frac{n_{ij}}{n} = \frac{n_{ij}}{n_{i.}} \times \frac{n_{i.}}{n} = f_{j/i} \times f_{i.}$$

ou $f_{ij} = \frac{n_{ij}}{n} = \frac{n_{ij}}{n_{.j}} \times \frac{n_{.j}}{n} = f_{i/j} \times f_{.j}$.

Les différentes fréquences sont liées par les relations suivantes :

$$f_{ij} = f_{j/i} \cdot f_{i.} = f_{i/j} \cdot f_{.j} \cdot$$

Les caractéristiques marginales et conditionnelles

Le tableau permet de calculer les caractéristiques habituelles des distributions.

– **Les caractéristiques marginales**

1. Pour la variable x :
   – la moyenne marginale

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{m} n_{i.} x_i = \sum_{i=1}^{m} f_{i.} x_i \ .$$

   – la variance marginale

$$V(x) = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{m}} \text{n}_{\text{i.}}(x_i - \overline{x})^2 = \frac{1}{\text{n}}\sum_{\text{i}=1}^{\text{m}} \text{n}_{\text{i.}} x_i^2 - \overline{x}^2 \ ,$$

$$V(x) = \sum_{\text{i}=1}^{\text{m}} \text{f}_{\text{i.}}(x_i - \overline{x})^2 = \sum_{\text{i}=1}^{\text{m}} \text{f}_{\text{i.}} x_i^2 - \overline{x}^2 \ .$$

2. Pour la variable y :
   – la moyenne marginale,

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j = \sum_{j=1}^{p} f_{.j} y_j \ ;$$

   – la variance marginale,

$$V(y) = \frac{1}{\text{n}}\sum_{j=1}^{p} n_{.j}(y_j - \overline{y})^2 = \frac{1}{n}\sum_{j=1}^{p} n_{.j} y_j^2 - \overline{y}^2$$

$$V(y) = \sum_{j=1}^{p} f_{.j}(y_j - \overline{y})^2 = \sum_{j=1}^{p} f_{.j} y_j^2 - \overline{y}^2 \ .$$

– **Les caractéristiques conditionnelles**

3. Pour la variable x :
   – les moyennes conditionnelles,

$$\overline{x}_j = \frac{1}{n_{.j}}\sum_{i=1}^{m} n_{ij} x_i = \sum_{i=1}^{m} f_{i/j} x_i \ ;$$

   – les variances conditionnelles,

$$V_j(x) = \frac{1}{n_{.j}}\sum_{i=1}^{m} n_{ij}(x_i - \overline{x}_j)^2 = \sum_{i=1}^{m} f_{i/j}(x_i - \overline{x}_j)^2 = \sum_{i=1}^{m} f_{i/j} x_i^{\ 2} - \overline{x}_j^{\ 2} \ .$$

4. Pour la variable y :
   – les moyennes conditionnelles,

$$\overline{y}_i = \frac{1}{n_{i\cdot}}\sum_{j=1}^{p} n_{ij} y_j = \sum_{j=1}^{p} f_{i/j} y_j \ ;$$

   – les variances conditionnelles,

$$V_i(y) = \frac{1}{n_{.j}}\sum_{j=1}^{p} n_{ij}(y_j - \overline{y}_i)^2 = \sum_{j=1}^{p} f_{i/j}(y_j - \overline{y}_i)^2 = \sum_{j=1}^{p} f_{i/j} y_j^2 - \overline{y}_i^2 \ .$$

Calculs pratiques dans un tableau de contingence

Dans le cas d'un tableau de contingence, il est souvent pratique d'utiliser le tableau suivant, surtout en s'aidant d'un tableur :

Tableau 7. Principe des calculs dans un tableau de contingence.

	y_j	Total	B_i	D_i	$\overline{y}_i$	$V_i(y)$
x_i	n_{ij}	$n_{i.}$	$B_i = \sum_{j=1}^{p} n_{ij} y_j$	$D_i = \sum_{j=1}^{p} n_{ij} y_j^2$	$\frac{B_i}{n_{i.}}$	$\frac{D_i}{n_{i.}} - \frac{B_i^2}{n_{i.}^2}$
Total	$n_{.j}$	n	B	D	$\overline{y} = \frac{B}{n}$	$V(y) = \frac{D}{n} - \frac{B^2}{n^2}$
A_j	$A_j = \sum_{i=1}^{m} n_{ij} x_i$	A				
C_j	$C_j = \sum_{i=1}^{m} n_{ij} x_i^2$	C				
$\overline{x}_j$	$\overline{x}_j = \frac{A_j}{n_j}$	$\overline{x} = \frac{A}{n}$				
$V_j(x)$	$\frac{C_j}{n_j} - \frac{A_j^2}{n_j^2}$	$V(x) = \frac{C}{n} - \frac{A^2}{n^2}$				

Il sera alors facile de calculer les caractéristiques des distributions en utilisant grandeurs intermédiaires calculées à l'aide du tableau ci-dessus.

Les caractéristiques conditionnelles de X :

$$\overline{x}_j = \frac{A_j}{n_{.j}}$$

$$V_j(X) = \frac{C_j}{n_{.j}} - \frac{A_j^2}{n_{.j}^2}.$$

Les caractéristiques marginales de X :

$$\overline{x} = \frac{A}{n}$$

$$V(x) = \frac{C}{n} - \frac{A^2}{n^2}.$$

Les caractéristiques conditionnelles de Y :

$$\overline{y}_i = \frac{B_i}{n_{i.}}$$

$$V_i(Y) = \frac{D_i}{n_{i.}} - \frac{B_i^2}{n_{i.}^2}.$$

Les caractéristiques marginales de Y :

$$\overline{y} = \frac{B}{n}$$

$$V(y) = \frac{D}{n} - \frac{B^2}{n^2}.$$

Exemple : le patrimoine mobilier et immobilier

Le tableau ci-dessous donne la répartition du patrimoine mobilier en fonction du patrimoine immobilier (évalué en unités d'habitation) pour l'ensemble des souscripteurs à une société d'assurances.

Calculez les caractéristiques marginales suivantes :

- la moyenne et la variance du patrimoine mobilier
- la moyenne et la variance du patrimoine immobilier
- Calculez des caractéristiques conditionnelles suivantes :
- la moyenne et la variance du patrimoine mobilier pour les sociétaires possédant une unité d'habitation.
- la moyenne et la variance des unités d'habitation patrimoine immobilier pour les sociétaires entre 200 et 300 milliers d'euros de patrimoine mobilier.

Tableau 8. Répartition des sociétaires assurés.

	Répartition du patrimoine mobilier (en milliers d'euros)						
UH	[0 ; 50[	[50 ; 100[	[100 ; 200[	[200 ; 300[	[300 ; 400[	[400 ; 500[	[500 ; 600]
c_i	25	75	150	250	350	450	550
0	870	970	650	170	50	10	5
1	180	950	1 820	840	330	170	45
2	30	150	450	340	170	110	35
3	10	30	80	80	50	40	15
4	5	10	50	40	30	20	10
5	1	5	30	20	20	10	10

Tableau 9. Les caractéristiques marginales de la distribution.

	25	75	150	250	350	450	550	$n_{i.}$	B_i	D_i	$\overline{y}_i$	$V_i(y)$
0	870	970	650	170	50	10	5	2 725	259 250	40 912 500	95,14	5 962,60
1	180	950	1 820	840	330	170	45	4 335	775 500	187 368 750	178,89	11 219,71
2	30	150	450	340	170	110	35	1 285	292 750	85 925 000	227,82	14 965,29
3	10	30	80	80	50	40	15	305	78 250	25 737 500	256,56	18 563,56
4	5	10	50	40	30	20	10	165	43 375	14 434 375	262,88	18 375,80
5	1	5	30	20	20	10	10	96	26 900	9 453 750	280,21	19 959,85
$n_{.j}$	1 096	2 115	3 080	1 490	650	360	120	8 911	1 476 025	363 831 875	165,64	89 046,81
A_j	295	1 405	3 310	2 020	1 040	640	250	8 960				
C_j	495	2 105	5 890	4 060	2 440	1 540	730	17 260				
$\overline{x}_j$	0,27	0,66	1,07	1,36	1,60	1,78	2,08	1,01				
$V_j(x)$	0,379	0,554	0,757	0,887	1,194	1,117	1,743	0,926				

Résultats

Les caractéristiques marginales sont obtenues directement par la lecture du tableau.

La moyenne du patrimoine mobilier :

$$\overline{c} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{147602}{8\,911} = 165{,}64 \text{ milliers d'euros.}$$

La variance s'obtient par : $V(c) = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \overline{c}^2 = \frac{363831875}{8\,911} - (165{,}64)^2 = 89046{,}81$

$$V(c) = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \overline{c}^2 = \frac{363831875}{8\,911} - (165{,}64)^2 = 89046{,}81 .$$

L'écart type est $\sigma_c = \sqrt{V(c)} = \sqrt{89046{,}81} = 298{,}4$ milliers d'euros.

La moyenne est : $\overline{u} = \frac{1}{n}\sum_{i=1}^{k} n_i u_i = \frac{8960}{8\,911} = 1{,}01$ unité d'habitation.

La variance s'obtient par : $V(u) = \frac{1}{n}\sum_{i=1}^{k} n_i u_i^2 - \overline{u}^2 = \frac{17260}{8911} - (1{,}01)^2 = 0{,}926$.

L'écart type est $\sigma_u = \sqrt{V(u)} = \sqrt{0{,}926} = 0{,}96$ unité d'habitation.

En ce qui concerne les caractéristiques conditionnelles de la distribution, nous arrivons aux résultats suivants.

Les caractéristiques statistiques pour les sociétaires possédant une unité d'habitation, ne sont autres que la distribution conditionnelle pour i = 2.

La moyenne est de : $\overline{c}_2 = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \frac{775500}{4\ 335} = 178{,}89$ milliers d'euros.

La variance s'obtient par :

$$V(c_2) = \frac{1}{n}\sum_{i=1}^{k} n_i c_i^2 - \overline{c}_2^2 = \frac{1873668750}{4\ 335} - (178{,}89)^2 = 11219{,}71$$

L'écart type est alors : $\sigma_{c_2} = \sqrt{V(c_2)} = \sqrt{11219{,}71} = 105{,}9$ milliers d'euros.

Pour les sociétaires ayant entre 200 et 300 milliers de euros de patrimoine mobilier, c'est la distribution conditionnelle pour j = 4.

$\overline{u}_4 = \frac{1}{n}\sum_{i=1}^{k} n_i u_i = \frac{2020}{1490} = 1{,}36$ unité d'habitation.

La variance s'obtient par :

$$V(u_4) = \frac{1}{n}\sum_{i=1}^{k} n_i u_i^2 - \overline{u}_4^2 = \frac{4060}{1490} - (1{,}36)^2 = 0{,}887 .$$

L'écart type est : $\sigma_{u_4} = \sqrt{V(u_4)} = \sqrt{0{,}887} = 0{,}94$ unité d'habitation.

La covariance

En statistique, la covariance de deux variables X et Y est la valeur :

$$\sigma_{xy} = \text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \overline{x})(y_i - \overline{y})$$

$$\sigma_{xy} = \text{cov}(x,y) = \sum_{i=1}^{n} f_i (x_i - \overline{x})(y_i - \overline{y})$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \text{et} \quad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2$$

Dans le cas d'un tableau de contingence, la formule de définition de la covariance est la suivante :

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{p} n_{ij} (x_i - \overline{x})(y_j - \overline{y})$$

ou $Cov(x,y) = \sum_{i=1}^{m}\sum_{j=1}^{p} f_{ij} (x_i - \overline{x})(y_j - \overline{y})$

Pour calculer les covariances, il est plus pratique d'utiliser les formules développées.

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i y_i - \overline{xy}$$

$$\sigma_{xy} = \mathrm{cov}(x,y) = \sum_{i=1}^{n} f_i x_i y_i - \overline{xy}$$

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} x_i y_j - \bar{x}\,\bar{y}$$

$$Cov(x,y) = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} x_i y_j - \bar{x}\,\bar{y}$$

Dans le cas d'un tableau de contingence, il est possible d'utiliser des résultats intermédiaires.

Tableau 10. Calcul de la covariance.

	y_j	**Total**	B_i	D_i	E_i
x_i	n_{ij}	$n_{i.}$	$B_i = \sum_{j=1}^{p} n_{ij} y_j$	$D_i = \sum_{j=1}^{p} n_{ij} y_j^2$	$E_i = x_i B_i$
Total	$n_{.j}$	n	B	D	E
A_j	$A_j = \sum_{i=1}^{m} n_{ij} x_i$	A			
C_j	$C_j = \sum_{i=1}^{m} n_{ij} x_i^2$	C			
E_j	$E_j = y_j A_j$	E			

$$Cov(x,y) = \frac{E}{n} - \frac{A}{n} \cdot \frac{B}{n}$$

Intuitivement, la covariance est une mesure de la variation simultanée de deux variables. La valeur absolue de la covariance augmente si les deux variables évoluent dans le même sens (covariance positive) ou en sens inverse (covariance négative). La notion de covariance pour les distributions à deux dimensions est analogue à celle de la variance pour les distributions à une dimension.

Exemple : Calcul d'une covariance dans un tableau de contingence

Tableau 11. Répartition du patrimoine mobilier (en milliers d'euros).

	25	75	150	250	350	450	550	$n_{i.}$	B_i	E_i
0	867	969	656	167	49	12	6	2 726	260 350	0
1	178	952	1 825	841	328	168	45	4 337	775 000	775 000
2	26	148	464	339	173	110	35	1 295	295 400	590 800
3	4	28	80	79	51	38	13	293	76 050	228 150
4	2	12	44	40	27	21	10	156	41 950	167 800
5	1	6	25	22	15	15	9	93	26 675	133 375
$n_{.j}$	1 078	2 115	3 094	1 488	643	364	118	**8 900**	**1 475 425**	**1 895 125**
A_j	255	1 410	3 294	2 026	1 010	661	239	**8 895**		
E_j	6 375	105 750	494 100	506 500	353 500	297 450	131 450	**1 895 125**		

La covariance s'exprime donc comme suit par application de la formule ci-dessus. Les données utilisées apparaissent en gras sur le tableau.

$$Cov(x,y) = \frac{E}{n} - \frac{A}{n} \cdot \frac{B}{n} = \frac{1895125}{8900} - \frac{8895}{8900} \cdot \frac{1475425}{8900} = 47{,}25$$

La recherche et l'estimation des liaisons

Il existe nombre de méthodes permettant de mettre en lumière les relations entre deux caractères statistiques. Nous ne présenterons ici que les plus usuelles, des indices de dépendance, pour des variables ordonnées nous avons le coefficient de Spearman, le Khi2 pour un tableau de contingence et l'ajustement linéaire ainsi que la corrélation pour des variables.

Les indicateurs de dépendance

Disposant de deux variables économiques, il est souvent pertinent de tester si elles sont indépendantes. L'hypothèse d'indépendance suppose que la connaissance d'une valeur d'une variable ne permet pas de disposer, de calculer, de prévoir la valeur de l'autre.
Les indicateurs de dépendance sont nombreux, leur objectif est identique, il s'agit de tester ou de montrer que les variations de variables, nous nous attarderons sur seulement le cas de deux variables dans cet ouvrage, dépendent d'une autre variable. Si deux variables dépendent l'une de l'autre, deux conséquences en découlent. Tout d'abord, il devient légitime de développer les analyses économiques qui permettent d'expliquer cette relation même si une réponse négative est tout aussi intéressante en ce sens qu'elle permet d'éliminer une hypothèse d'interaction. Ensuite, la connaissance des variations d'une variable permettra d'inférer les variations de l'autre, il devient possible de faire des conjectures voire des prévisions sur les valeurs de la seconde.

Les indicateurs présentés permettent d'appréhender les techniques utilisées sans aucune prétention d'exhaustivité. Le plus simple ne retient que le sens des évolutions, les plus complexes retiennent la valeur des évolutions ou des classements.

L'indice de dépendance

Il rend compte de la proportion de variations concordantes et discordantes entre les deux variables X et Y. Il y a concordance quand les deux variables évoluent dans le même sens et discordance dans le cas inverse. Cet indice sera d'autant plus près de 1 en valeur absolue que la relation sera forte, son signe nous indique si les deux variables varient dans le même sens. Il ne retient que le signe des variations sans les pondérer par leur importance.
L'indice de dépendance I_D est égal à :

$$I_D = \frac{nombre\ de\ concordances - nombres\ de\ discordances}{nombre\ total}$$

Exemple

Tableau 12. Indice de dépendance.

Périodes	X	Variations	Y	Variations	Produits
1	50		100		
2	52	+	120	+	+
3	60	+	122	+	+
4	62	+	128	+	+
5	72	+	125	-	-
6	80	+	124	-	-

La colonne « produits » indique trois concordances (signe +) et deux discordances (signe -), l'indice est de :

$$I_D = \frac{3-2}{5} = 0{,}2 \cdot$$

On peut donc en conclure que les deux variables ne sont que faiblement dépendantes l'une de l'autre.

Le coefficient de dépendance

Le coefficient de dépendance tient compte de l'importance des écarts. Les concordances et les discordances sont estimées par le produit algébrique des différences. La valeur absolue du coefficient signale l'intensité de la liaison, son signe indique dans quelle mesure les deux variables évoluent l'une par rapport à l'autre.

Tableau 13. Coefficient de dépendance.

Périodes	X	Variations	Y	Variations	Produit
1	50		100		
2	52	+ 2	120	+ 20	+ 40
3	60	+ 8	122	+ 2	+ 16
4	62	+ 2	128	+ 6	+ 12
5	72	+ 10	125	- 3	- 30
6	80	+ 8	124	- 1	- 8

C = somme des produits des convergences = 68

D = somme des produits des divergences = 38

$$C_D = \frac{C - D}{C + D} = \frac{68 - 38}{106} = \frac{30}{106} = 0{,}283$$

Le coefficient de dépendance confirme la faible relation entre les deux variables mises en lumière par l'indice de dépendance.

– **Le coefficient de corrélation des rangs de Spearman**

Le coefficient des rangs de Spearman[1] sert à déterminer la relation qui existe entre deux séries de données. Dans le cas de deux variables classées – par ordre de taille ou d'importance –, le coefficient de Spearman fait intervenir uniquement les notions de rang. Il mesure la relation qui existe entre deux classements. Ce coefficient peut être calculé, que la variable soit quantitative ou qualitative à condition qu'elle soit ordinale – on ne peut classer les individus selon les pays de naissance par exemple. Il mesure le degré de la relation entre les classements des informations selon la variable X et le classement des données selon la variable Y. On calcule le rang de chaque élément dans la série croissante de X et de Y, puis on calcule la différence de classement di pour chaque couple de valeur.

Il est calculé par la formule suivante :

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

avec $d_i = |X_i - Y_i|$ où X_i est le rang de l'item i dans un premier classement et Y_i son rang dans un second. Dans la formule n correspond au nombre de couples de valeur (X, Y).

Plus, le coefficient est proche de 1, plus la corrélation de rang est forte.

- $\rho \cong 1$ il y a concordance
- $\rho \cong -1$ il y a discordance (les deux variables varient en sens inverse)
- $\rho \cong 0$ pas de relation.

Les controverses sont nombreuses quant à la relation entre le taux de croissance du PIB et celui de la hausse des prix, c'est-à-dire le taux d'inflation. Les données suivantes donnent pour 2013 les deux taux pour les pays membres de l'Union européenne.

1. Le psychologue Charles Spearman introduit ce coefficient en 1904.

Tableau 14. Coefficient de Spearman.

États membres	Taux de croissance du PIB	États membres	Taux d'inflation
Allemagne	0,4	Allemagne	1,6
Autriche	0,3	Autriche	2,1
Belgique	0,2	Belgique	1,2
Bulgarie	0,9	Bulgarie	0,4
Chypre	-5,4	Chypre	0,4
Croatie	-0,9	Croatie	2,3
Danemark	0,4	Danemark	0,5
Espagne	-1,2	Espagne	1,5
Estonie	0,8	Estonie	3,2
Finlande	-1,4	Finlande	2,2
France	0,2	France	1
Grèce	-3,9	Grèce	-0,9
Hongrie	1,1	Hongrie	1,7
Irlande	-0,3	Irlande	0,5
Italie	-1,9	Italie	1,3
Lettonie	4,1	Lettonie	0
Lituanie	3,3	Lituanie	1,2
Luxembourg	2,1	Luxembourg	1,7
Malte	2,6	Malte	1
Pays-Bas	-0,8	Pays-Bas	2,6
Pologne	1,6	Pologne	0,8
Portugal	-1,4	Portugal	0,4
République tchèque	-0,9	République tchèque	1,4
Roumanie	3,5	Roumanie	3,2
Royaume-Uni	1,7	Royaume-Uni	2,6
Slovaquie	0,9	Slovaquie	1,5
Slovénie	-1,1	Slovénie	1,9
Suède	1,6	Suède	0,4

Source des données : Eurostat

États membres	Classements		*D*	*D*[2]
	Taux de croissance	Taux inflation		
Lettonie	1	27	-26	676
Roumanie	2	2	0	0
Lituanie	3	17	-14	196
Malte	4	19	-15	225
Luxembourg	5	9	-4	16
Royaume-Uni	6	4	2	4
Pologne	7	20	-13	169
Suède	8	26	-18	324
Hongrie	9	10	-1	1
Bulgarie	10	24	-14	196
Slovaquie	11	13	-2	4
Estonie	12	1	11	121
Danemark	13	21	-8	64
Allemagne	14	11	3	9
Autriche	15	7	8	64
Belgique	16	16	0	0
France	17	18	-1	1
Irlande	18	22	-4	16
Pays-Bas	19	3	16	256
République tchèque	20	14	6	36
Croatie	21	5	16	256
Slovénie	22	8	14	196
Espagne	23	12	11	121
Portugal	24	25	-1	1
Finlande	25	6	19	361
Italie	26	15	11	121
Grèce	27	28	-1	1
Chypre	28	24	4	16
:				3 451

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 3451}{28(28^2-1)} \cong 0{,}056$$

Le lien entre les deux indicateurs selon le coefficient de Spearman est statistiquement très faible.

Indépendance dans un tableau de contingence

Deux variables seront indépendantes si les distributions conditionnelles de chacune des deux variables sont identiques à la distribution marginale et donc identiques entre elles. Dans ce cas les moyennes conditionnelles seront égales entre elles et égales à la moyenne marginale. Une variable sera dite indépendante d'une autre si les répartitions conditionnelles de cette variable sont identiques à sa répartition marginale.
En termes de fréquence, cela signifie que $f_{i/j} = f_{i.}$ ou, symétriquement, que $f_{j/i} = f_{.j}$. Autrement dit que, quelle que soit la valeur prise par j la fréquence de i est toujours la même.

$f_{j/i} = \frac{f_{il}}{f_{;j}} = f_{i.}$ donc $f_{ij} = f_{i.} \cdot f_{.j}$

On dit que les variables X et Y sont totalement indépendantes si les variations de l'une n'entraînent pas de variations de l'autre. Autrement dit, la variable Y est indépendante de la variable X si les fréquences conditionnelles $f_{j/i}$ sont égales entre elles pour j fixé c'est-à-dire, ne dépendent pas de i.
Une conséquence immédiate est que les fréquences conditionnelles sont égales aux fréquences marginales en cas d'indépendance totale.

$f_{j/i} = f_{.j}$

En considérant par exemple la répartition des salariés d'une entreprise selon le montant du salaire et l'âge : le salaire est indépendant de l'âge si, parmi les salariés des différentes tranches d'âge, la proportion de ceux dont le salaire est compris entre telle et telle limite ne varie pas d'une tranche d'âge à l'autre. Ceci revient à dire que les lignes du tableau de contingence donnant les effectifs n_{ij} sont proportionnelles entre elles et donc proportionnelles à la ligne marginale.
La relation $f_{ij} = f_{i.} f_{j/i} = f_{.j} f_{i/j}$; indique que si Y est indépendante de X on a : $f_{i.} f_{.j} = f_{.j} f_{j/i} \Rightarrow f_{i/j} = f_{i.}$ et donc que X est indépendante de Y, autrement dit, l'indépendance est réciproque.
On en déduit alors : $f_{ij} = f_{i.} \cdot f_{.j}$ ou encore : $\frac{n_{ij}}{n} = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}$.
En résumé, le tableau de contingence associé aux deux caractères X et Y indique que X et Y sont indépendants si on a la relation suivante entre les effectifs :

$n \cdot n_{ij} = n_{i.} \cdot n_{.j}$

$n_{ij} = \frac{n_{i.} n_{.j}}{f_{;j}}$

Le khi deux

Dans un tableau de contingence, la comparaison des distributions conditionnelles et marginales permet de tester la dépendance entre deux variables.

Deux variables seront indépendantes si les distributions conditionnelles de chacune des deux variables sont identiques à la distribution marginale et donc identiques entre elles. En termes de fréquence, cela signifie que $f_{i/j} = f_{i.}$ ou symétriquement que $f_{j/i} = f_{.j}$. Autrement dit que quelle que soit la valeur prise par j la fréquence de i est toujours la même ; donc que $f_{j/i} = f_{i.}f_{.j}$ ou que $n \cdot n_{j/i} = n_{i.}n_{.j}$.

Le test du khi2 qui s'écrit χ^2 est utilisé pour estimer l'indépendance au sens défini ci-dessus. C'est un indicateur d'écart entre la série empirique, celle dont nous disposons, et une série théorique construite sur des effectifs conjoints égaux au produit des produits des fréquences marginales par l'effectif total. Les effectifs théoriques correspondent à l'hypothèse d'une totale indépendance des variables ou caractères.

La formule générale du χ^2 est la suivante :

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(n_i - n_i^*\right)^2}{n_i^*}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(n_i - n_i^*\right)^2}{n_i^*} = \sum_{i=1}^{n} \frac{n_i^2 - 2n_i n_i^* + n_i^{*22}}{n_i^*} = \sum_{i=1}^{n} \frac{n_i^{22}}{n_i^*} - 2\sum_{i=1}^{n} n_i + \sum_{i=}^{n} n_i^*$$

$$\chi^2 = \sum_{i=1}^{n} \frac{n_i^{22}}{n_i^*} - 2n + n = \sum_{i=1}^{n} \frac{n_i^{22}}{n_i^*} - n$$

avec n_i, l'effectif observé et n_i^* l'effectif théorique.
Nous pouvons écrire le calcul du χ^2 en utilisant une autre formalisation pour un tableau de contingence :

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{n} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*} \;;\; \chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{n} \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

avec n_{ii} l'effectif observé et n_{ii}^* l'effectif théorique, f_{ij} la fréquence observée, $f_{ij}^* = f_{i.} \cdot f_{;j}$ la fréquence théorique.

Tableau 15. Série théorique.

Variable 2 / Variable 1	Modalité 1	Modalité j	Modalité p	Effectifs marginaux de la variable 1
Modalité 1	$n_{11}^* = \frac{n_{1.} \cdot n_{.1}}{n}$	$n_{1j}^* = \frac{n_{1.} \cdot n_{.j}}{n}$	$n_{1p}^* = \frac{n_{1.} \cdot n_{.p}}{n}$	$n_{1.}$
Modalité i	$n_{i1}^* = \frac{n_{i.} \cdot n_{.1}}{n}$	$n_{ij}^* = \frac{n_{i.} \cdot n_{.j}}{n}$	$n_{ip}^* = \frac{n_{i.} \cdot n_{.p}}{n}$	$n_{i.}$
Modalité m	$n_{m1}^* = \frac{n_{m.} \cdot n_{.1}}{n}$	$n_{mj}^* = \frac{n_{m.} \cdot n_{.j}}{n}$	$n_{mp}^* = \frac{n_{m.} \cdot n_{.p}}{n}$	$n_{m.}$
Effectifs marginaux de la variable 2	$n_{.1}$	$n_{.j}$	$n_{.p}$	n

Il est possible aussi d'apprécier le niveau de signification de la liaison en utilisant une « Table du χ^2 » ; c'est-à-dire, une table donnant les valeurs critiques permettant d'accepter la dépendance entre les variables.
Si le χ^2 a une valeur élevée, c'est l'indice que le hasard seul est insuffisant pour expliquer la valeur trouvée, donc une liaison est vraisemblable. Inversement, si le χ^2 est faible alors, le hasard suffit à expliquer les différences entre le tableau observé et le tableau théorique d'indépendance.
Un exemple illustre l'intérêt du χ^2, pour deux caractères qualitatifs.

Exemple : Les inscriptions de bacheliers dans l'enseignement supérieur selon les PCS

Tableau 16. Origine des bacheliers s'inscrivant dans l'enseignement supérieur en 2005.

Effectifs constatés	Université	CPGE (1)	STS (2)	Ensemble
Agriculteurs	5 105	757	2 831	8 694
Chefs d'entreprise	17 991	3 181	9 514	30 685
Cadres	73 422	19 578	15 743	108 742
Professions intermédiaires	39 628	5 339	18 461	63 429
Employés	38 899	3 257	19 594	61 749
Ouvriers	32 335	1 856	25 710	59 900
Retraités	22 124	2 802	15 516	40 442
Indéterminé	13 615	1 098	5 889	20 602
	243 118	37 868	113 258	394 244

(1) CPGE : classes préparatoires aux grandes écoles
(2) STS : Sections de techniciens supérieurs
Source : Ministère de l'Éducation nationale – DEPP

La formule de l'effectif théorique est $n_{ij}^{*} = \frac{n_{i.} \cdot n_{.j}}{n}$. Pour calculer l'effectif théorique des enfants d'agriculteurs en CPEG le calcul est le suivant (les résultats sont arrondis aux entiers) :

$$n_{agriCPEG}^{*} = \frac{n_{i.} \cdot n_{.j}}{n} = \frac{8694 \cdot 243118}{394244} \cong 835 .$$

Tableau 17. Effectifs théoriques.

Effectifs théoriques	Université	CPGE	STS	Ensemble
Agriculteurs	5 361	835	2 498	8 694
Chefs d'entreprise	18 922	2 947	8 815	30 685
Cadres	67 058	10 445	31 239	108 742
Professions intermédiaires	39 115	6 092	18 222	63 429
Employés	38 079	5 931	17 739	61 749
Ouvriers	36 938	5 754	17 208	59 900
Retraités	24 939	3 885	11 618	40 442
Indéterminé	12 705	1 979	5 919	20 602
	243 117	37 868	113 258	394 243

L'écart d'une unité s'explique par les arrondis

Calcul du $\chi^2 = \frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}}$

Tableau 18. Calcul du χ^2.

Khi 2	Université	CPGE	STS	Ensemble
Agriculteurs	12,3	7,3	44,5	64,1
Chefs d'entreprise	45,9	18,5	55,4	119,8
Cadres	604,0	7 986,0	7 687,0	16 277,0
Professions intermédiaires	6,7	93,2	3,1	103,1
Employés	17,7	1 205,7	193,9	1 417,3
Ouvriers	573,7	2 640,2	4 200,6	7 414,5
Retraités	317,8	301,7	1 307,7	1 927,2
Indéterminé	65,2	392,1	0,1	457,5
Ensemble	1 643,3	12 644,7	13 492,4	27 780,4

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{n} \frac{\left(n_{ij} - n_{ij}^{*}\right)^2}{n_{ij}^{*}} = 27780$$

Le χ^2 est suffisamment important pour que nous puissions rejeter l'hypothèse d'une indépendance des deux variables. L'utilisation de la table nécessite de déterminer le nombre de degrés de liberté du tableau, noté v nombre de degrés de liberté. Ce nombre est $v = (p-1)(q-1)$.
La table du khi-deux χ^2 ci-dessous donne les valeurs critiques permettant de rejeter l'hypothèse d'indépendance des deux variables avec un risque d'erreur de 10 % et de 5 %.

Tableau 19. Extrait de la table du modèle de χ^2 : (Test de Pearson).

v	**0.1**	**0.05**
1	2.71	3.84
2	4.61	5.99
3	6.25	7.81
4	7.78	9.49
5	9.24	11.07
6	10.64	12.59
7	12.02	14.07
8	13.36	15.51
9	14.68	16.92
10	15.99	18.31
11	17.27	19.67
12	18.55	21.03
13	19.81	22.36
14	21.06	23.68
15	22.31	25.00

Le nombre de degrés de liberté v est de 14 ($v = (3-1)(8-1)$), la valeur critique du χ^2 est au seuil de 5 % de 23,68. Les inscriptions des nouveaux étudiants dans les filières de l'enseignement supérieur dépendent de la catégorie sociale des parents. Toutefois, ce n'est pas le résultat du χ^2 qui permet cette affirmation car le χ^2 est une analyse purement descriptive et symétrique, c'est uniquement la connaissance du contexte qui laisse penser que la CSP influe sur le choix des études et non le choix des études qui explique la CSP.
L'examen de la table des contributions au χ^2 permet de repérer les cellules les plus significatives.

$$contribution_{ij} = \frac{\frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}}{\sum_{ij} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}}.$$

Tableau 20. Tableau des contributions au Cχ^2.

Contributions	Université	CPGE	STS	Ensemble
Agriculteurs	0,0	0,0	0,2	0,2
Chefs d'entreprise	0,2	0,1	0,2	0,4
Cadres	2,2	**28,7**	**27,7**	58,6
Professions intermédiaires	0,0	0,3	0,0	0,4
Employés	0,1	4,3	0,7	5,1
Ouvriers	2,1	**9,5**	**15,1**	26,7
Retraités	1,1	1,1	4,7	6,9
Indéterminé	0,2	1,4	0,0	1,6
Ensemble	5,9	45,5	48,6	100,0

Aux arrondis près

À la lecture de ce tableau quatre pourcentages importants (supérieurs à 5) se détachent : la proportion des enfants de « Cadres » en CPGE en STS et celle des enfants d'« Ouvriers » en STS et secondairement en CPGE. Ils représentent 81 % du total du χ^2. Les autres contributions sont faibles. Ces résultats en partie contre-intuitifs méritaient une analyse détaillée.

Le V de Cramer

Indépendamment de l'utilisation des tables du χ^2, il convient en statistique descriptive d'apprécier la significativité de la liaison et donc la valeur du χ^2 calculé. Il est difficile de dire si la liaison est forte ou faible au seul vu d'un résultat. Il est nécessaire d'avoir un élément de comparaison.
Pour cela on calcule le rapport du χ^2 au χ^2 maximum que l'on aurait obtenu si la dépendance avait été totale.
On démontre que ce χ^2 maximum vaut : $\chi^2 \max = n \cdot Min(p-1;q-1)$.

Avec p le nombre de lignes du tableau et q le nombre de colonnes.

Le rapport des deux χ^2 est le V^2 de Cramer : $V^2 = \dfrac{\chi^2}{\chi^2 \max}$.

Ce nombre V^2 est un nombre compris entre 0 et 1. Plus il est proche de 1, plus la dépendance entre les variables est forte.

$$Min(p-1;q-1) = Min(7;2) = 2$$

$$\chi^2 \max = 394244 \cdot 2 = 788488$$

$$V^2 = \frac{\chi^2}{\chi^2 \max} = \frac{27780}{788488} = 0{,}035231988$$

$$V = \sqrt{0{,}035231988} = 0{,}18770186$$

Le V de Cramer indique que les CSP et les choix de formation pour les nouveaux bacheliers sont faiblement dépendants. Bien que l'analyse utilisant le test du χ^2 ait montré que les variables étaient significativement dépendantes, la faiblesse du V Cramer est susceptible de s'expliquer par le fait que 81 % des contributions sont concentrées dans quatre cellules du tableau.

L'ajustement linéaire

La mise en évidence d'une relation entre deux variables signifie que leurs évolutions statistiques sont liées. Dans ce contexte, l'ajustement estime une relation fonctionnelle entre les variables. L'équation servira à estimer la valeur prise par une des variables à partir de la valeur prise par l'autre. Dans le cadre de cet ouvrage, la présentation se limitera à la détermination des coefficients des droites d'ajustement

Si l'analyse économique théorique postule l'existence d'une liaison entre deux variables, la confirmation statistique se déroule en trois étapes.

- 1. C'est la recherche du modèle explicite, de la forme de la relation entre les deux variables, c'est-à-dire, celui qui précise formellement l'équation de la courbe adaptée à l'allure du phénomène.
- 2. Cette étape consiste à estimer les coefficients du modèle de la relation. Cette équation servira à estimer la valeur prise par une des variables en fonction de la valeur prise par l'autre.
- 3. Enfin, la troisième étape donne une mesure de l'intensité de la liaison postulée, de la qualité de l'ajustement.

Nous examinerons ces trois étapes dans le cas de l'ajustement linéaire.

La recherche de la forme de la relation

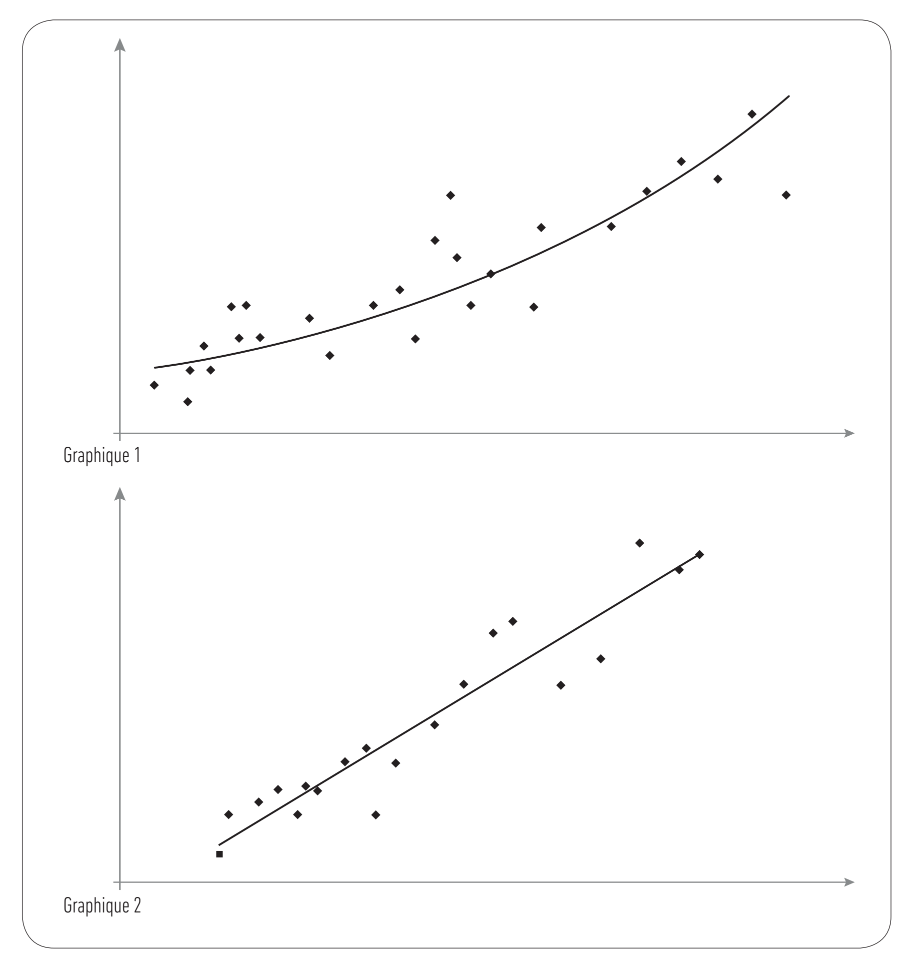
La forme de la liaison peut être postulée, c'est le calcul statistique qui en testera la validité. Dans le cas de deux variables, la méthode empirique la plus simple est de faire une représentation graphique.

La mise en lumière graphique d'une liaison

Dans le cas d'une distribution à deux caractères, la méthode d'ajustement la plus simple – ce qui ne veut pas dire ni la moins fiable ni la moins efficace – c'est l'ajustement graphique. Cette méthode consiste à tracer, à main levée, une courbe régulière aussi simple que possible et passant au travers des points représentatifs des observations, de façon à ce que les écarts positifs et négatifs se compensent.

Cette méthode peut paraître arbitraire mais l'expérience montre que les courbes tracées par différentes personnes sont proches les unes des autres. Sa précision reste limitée, elle ne fait pas illusion sur la précision des résultats. La représentation graphique met, souvent, en évidence la forme générale de la liaison et fournit une première appréciation de l'intensité de celle-ci, une dépendance statistique ou corrélation entre les variables observées.

Figure 3. Des liaisons possibles.



Les deux graphiques montrent deux courbes d'ajustement différentes, le premier une tendance exponentielle croissante, la seconde une tendance linéaire décroissante. La méthode graphique ne permet pas d'aller bien loin dans l'analyse, les méthodes analytiques permettent une approche plus fine et plus précise.

La méthode graphique ne permet pas d'aller bien loin dans l'analyse, les méthodes analytiques permettent une approche plus fine et plus précise.

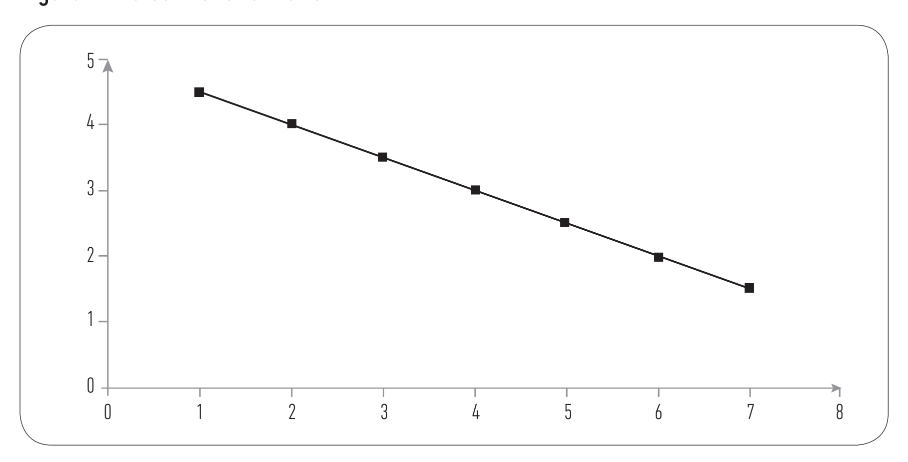
La signification des courbes d'ajustement

La courbe d'ajustement rend plus ou moins compte de l'ensemble des données.

– La liaison fonctionnelle

Dans ce cas du phénomène représenté par le graphique suivant, tous les points appartiennent à la courbe.

Figure 4. Liaison fonctionnelle.

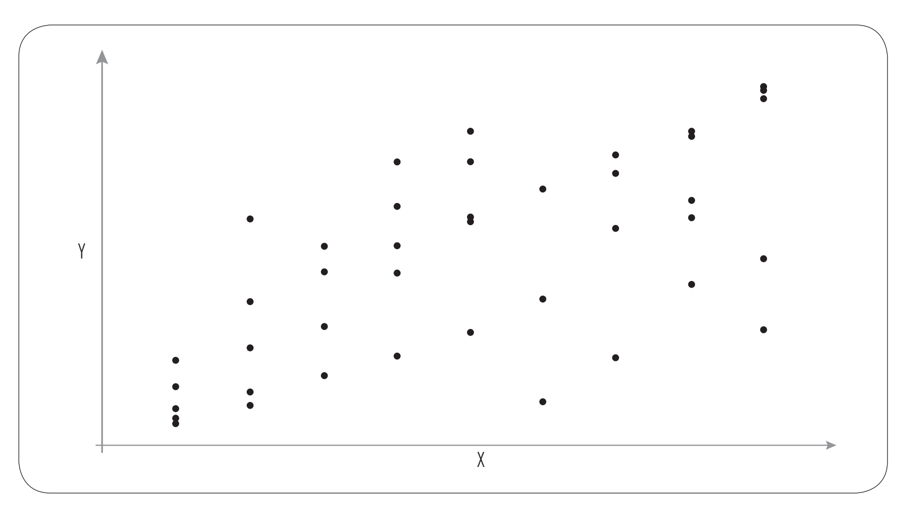


La liaison fonctionnelle réciproque entre Y et X entraîne qu'à chaque valeur x_i correspond une valeur et une seule de y_i et réciproquement. Le traitement de ce type de liaison ressort plus des mathématiques que de la statistique. La relation est du type $y = f(x)$.

– L'absence de relation

Un second cas échappe, par définition, à l'analyse statistique si les variables X et Y sont totalement indépendantes.

Figure 5. Aucune relation apparente.

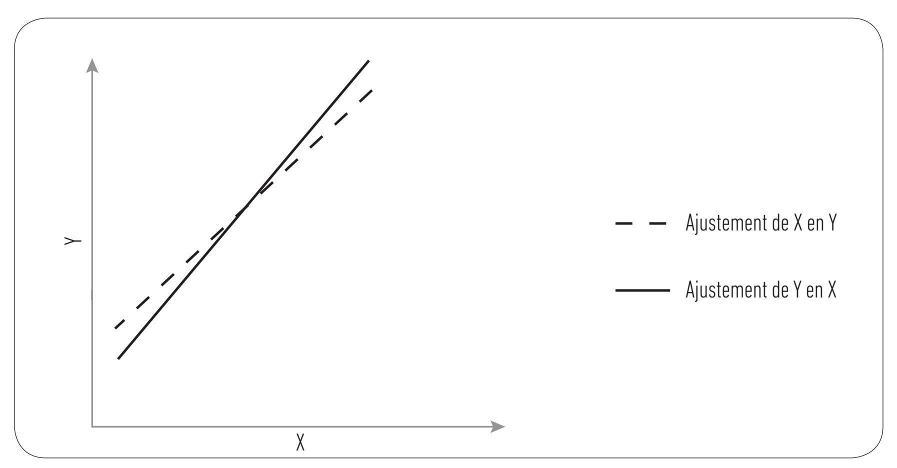


La connaissance de la valeur d'une variable n'apporte aucune information quant à la distribution de l'autre.

– La corrélation

Ce cas est celui où s'appliquent les méthodes statistiques. Il existe un lien entre les deux variables. La connaissance de la valeur prise par une variable permet une évaluation de la valeur de l'autre. Le modèle est explicatif et prévisionnel.

Figure 6. Les deux formes des courbes d'ajustement.



La connaissance de la valeur prise par X apporte une information supplémentaire sur les valeurs possibles de Y, on dit que Y est en corrélation avec X et que X est en corrélation avec Y car la relation est réciproque. Il existe

une certaine dépendance entre X et Y. La corrélation peut être directe ou positive quand les deux variables évoluent dans le même sens, si les variations sont de sens contraire ; on dit que la corrélation est inverse ou négative.

Figure 7. Corrélation positive.

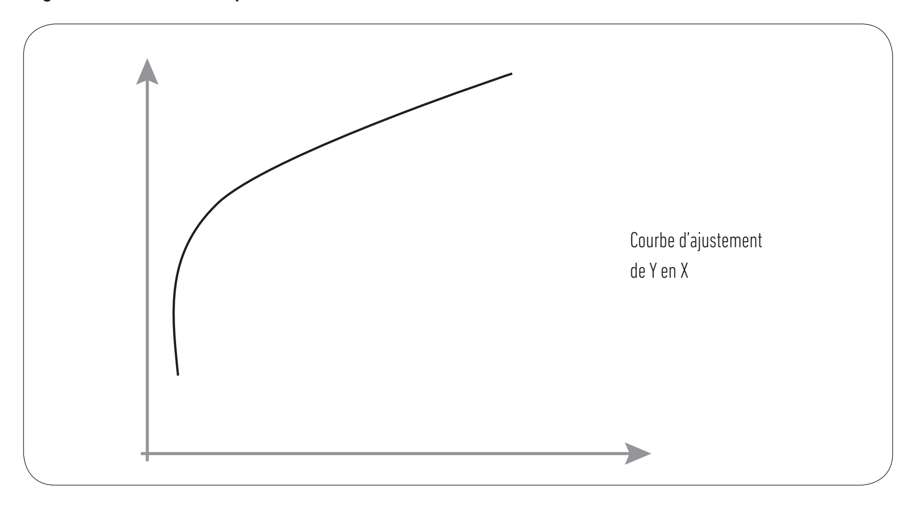
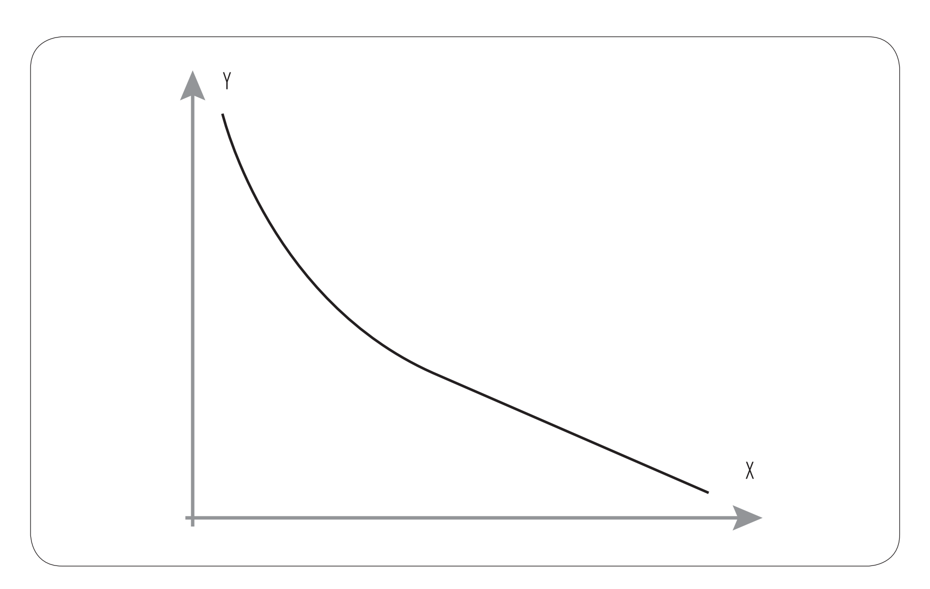


Figure 8. Corrélation négative.



Lorsque les courbes d'ajustement sont des droites non parallèles aux axes de coordonnées, il y a corrélation linéaire.

L'ajustement affine ou linéaire

Si l'étude graphique permet d'estimer que les points du diagramme suivent une droite, il est légitime de calculer l'équation de la droite.
Le but est d'obtenir un ajustement entre deux variables à partir des données disponibles et de déterminer les coefficients a et b d'une fonction affine : $f(x) = ax + b$.

La valeur yi de Y pour la valeur xi est égale à $y_i = f(x_i) + e_i$ où e_i est l'écart entre la valeur réelle yi et la valeur estimée $\hat{y}_i = f(x_i)$; et $e_i = y_i - \hat{y}_i$. Le problème est de déterminer la droite d'équation $y = f(x)$ satisfaisant au mieux $y_i = f(x_i)$.

Les données peuvent être individualisées ou regroupées en classe ; la différence est que, dans ce second cas, les évaluations obtenues seront plus incertaines. Deux méthodes seront présentées, une méthode empirique dite de Mayer et la méthode universellement utilisée des moindres carrés.

– La méthode de Mayer

C'est une méthode simple à mettre en œuvre proposée par Tobias Mayer, un astronome allemand du XVIIe siècle. La distribution est partagée en deux groupes comportant un nombre égal de données (effectif pair) ou un nombre égal de données à l'unité près (effectif total impair). En annulant la somme des écarts $\sum_{i=1}^{k} e_i = 0$ dans les deux groupes, on obtient deux équations permettant de déterminer a et b.

Soit la distribution suivante :

Tableau 21. Distribution des valeurs.

x_i	1	2	3	4	5
y_i	30	10	20	40	25

L'équation de la droite est de la forme suivante : $\hat{y}_i = ax_i + b$

La distribution est subdivisée en deux groupes dans la configuration ci-dessous.

Tableau 22. Système à résoudre.

Groupe 1 (trois données)	Groupe 2 (deux données)
e_1 = 30–a–b	e_4 = 40–4a–b
e_2 = 10–2a–b	e_5 = 25–5a–b
e_3 = 20–3a–b	

Une méthode de résolution est la suivante : $e_i = y_i - \hat{y}_i = y_i - ax_i + b$.
avec $\sum_{i=1}^{k} e_i = 0$ pour chaque groupe de données.

Pour le groupe 1 : $\sum_{i=1}^{3} e_i = 60 - 6a - 3b = 0$; $6a + 3b = 60$.

Pour le groupe 2 : $\sum_{i=1}^{2} e_i = 65 - 9a - 2b = 0$; $9a + 2b = 65$.

Il suffit de résoudre le système à deux équations et deux inconnues :
$6a + 3b = 60$;
$9a + 2b = 65$
d'où : a = 5 et b = 10.
$\hat{y}_i = 5x_i + 10$

Tableau 23. Système à résoudre second regroupement.

Groupe 1 (trois données)	Groupe 2 (deux données)
e_1 = 30–a–b	e_3 = 20–3a–b
e_2 = 10–2a–b	e_4 = 40–4a–b
	e_5 = 25–5a–b

Dans cette configuration a=50/15 et b=15.

Tableau 24. Tableau des résultats.

x_i	1	2	3	4	5
x_i	30	10	20	40	25
Configuration 1					
$\hat{y}_i$	15	20	25	30	35
e_1	15	-10	-5	10	-10
Configuration 2					
$\hat{y}_i$	18,33	21,67	25,00	28,33	31,67
e_1	11,67	-11,67	-5,00	11,67	-6,67

Les deux résultats sont assez différents, le choix arbitraire des regroupements ne permet pas une grande précision. L'estimation y_6 de pour $x_i = 6$ est de 40 dans la première configuration et de 35 dans la seconde.
L'ajustement sera d'autant meilleur que les sous-ensembles seront choisis judicieusement.

Cette méthode empirique est une première approche. Elle fournit des estimations numériques plus précises qu'une simple détermination graphique. La méthode des moindres carrés ordinaires donne une solution plus satisfaisante

– La méthode des moindres carrés ordinaires (MCO)

Les coefficients numériques d'une fonction affine $f(x) = ax + b$ doivent être tels que les écarts, $e_i = y_i - \hat{y}_i$, comptés parallèlement à l'axe des y, soient aussi faibles que possible. Une première hypothèse est de chercher les coefficients a et b tels que la somme des écarts soit nulle :

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - ax_i - b) = \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = n\overline{y} - an\overline{x} - nb$$

$$\overline{y} - a\overline{x} - b = 0$$

en définitive cela donne la formule : $\overline{y} = a\overline{x} + b$.
Cette condition signifie que toute droite passant par le point moyen vérifie la condition posée. Il existe alors une infinité de solutions, ce qui n'est pas satisfaisant.
Cela conduit à tester une nouvelle condition : minimiser la somme des carrés des écarts toujours parallèlement à l'axe des ordonnées. Cette méthode, dite des moindres carrés ordinaires (MCO) fournit une évaluation des coefficients de la droite telle que la somme des carrés des écarts des points à cette droite mesurée parallèlement à l'axe de la variable expliquée soit minimum. La droite satisfaisant à cette condition est celle qui présente le maximum de vraisemblance compatible avec le type de fonction choisie et avec l'ensemble des observations disponibles.
Quelles sont les conditions qui minimisent la somme des écarts ?

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left[y_i - f(x_i)\right]^2$$

avec $f(x_i) = \hat{y}_i = ax_i + b$, soit : $S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i - (ax_i + b)\right]^2$.

La condition première de minimisation de cette expression est que les dérivées premières en *a* et *b* soient nulles.
La somme S est minimum quand les dérivées partielles $\frac{\partial S}{\partial a}, \frac{\partial S}{\partial b}$ sont nulles simultanément.

Les deux dérivées sont :

$$\begin{cases} \dfrac{\partial \sum_{i=1}^{n} [y_i - (ax_i + b)]^2}{\partial a} = -2\sum_{i=1}^{n} [y_i - (ax_i + b)]x_i \\ \dfrac{\partial \sum_{i=1}^{n} [y_i - (ax_i + b)]^2}{\partial b} = -2\sum_{i=1}^{n} [y_i - (ax_i + b)] \end{cases}$$

Elles fournissent les conditions de minimisation (équations normales ou équations de Gauss) :

$$\begin{cases} (1)\ \sum_{i=1}^{n} [y_i - (ax_i + b)] \cdot x_i = 0 \\ (2)\ \sum_{i=1}^{n} [y_i - (ax_i + b)] = 0 \end{cases}$$

La condition (2) signifie que la droite d'ajustement de Y en X passe par le point moyen ($\bar{x}, \bar{y}$), elle est identique à la condition d'annulation de la somme des écarts : $b = \bar{y} - a\bar{x}$.

En reportant la valeur trouvée de **b** dans l'équation (1) il est possible d'obtenir a.

$$\sum_{i=1}^{n} \left[y_i - (ax_i + b) \right] x_i = \sum_{i=1}^{n} \left[y_i - (ax_i + \bar{y} - a\bar{x}) \right] x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 + a\bar{x}\sum_{i=1}^{n} x_i - \bar{y}\sum_{i=1}^{n} x_i = 0$$

$$a\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\ \bar{x}\ \bar{y}$$

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\ \bar{x}\ \bar{y}}{\sum_{i=1}^{n} \mathrm{x}_\mathrm{i}^2 - n\ \bar{\mathrm{x}}^2}$$

La pente de la droite d'ajustement de Y en X est ainsi déterminée.

Le numérateur et le dénominateur de cette formule ne sont que les expressions développées de la covariance de X et Y et de la variance de X. Le coefficient directeur de la droite d'ajustement **a** peut s'écrire comme le rapport de la covariance de Y et de la variance de X.

$$a = \frac{Cov(X,Y)}{V(X)}$$

$$\hat{y}_i = \frac{Cov(X,Y)}{V(X)} x_i + (\overline{y} - \frac{Cov(X,Y)}{V(X)} \overline{x})$$

$$(\hat{y}_i - \overline{y}) = \frac{Cov(X,Y)}{V(X)} (x_i - \overline{x})$$

Le choix de minimiser les écarts en fonction de la variable Y est une des possibilités. Il est envisageable de la même manière de chercher à minimiser les écarts parallèlement à l'axe des X, afin d'obtenir l'équation de la droite d'ajustement de X en Y. Cette droite aura pour équation $\hat{x} = a'y + b'$.
Ce qui revient à minimiser la somme S'

$$S' = \sum_{i=1}^{n} e_i'^2 = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2$$

La somme S'est minimum quand les dérivées partielles $\frac{\partial S'}{\partial a'}$ et $\frac{\partial S'}{\partial b'}$ sont nulles simultanément, d'où les deux conditions de Gauss :

$$\sum_{i=1}^{n} \left[x_i - (a'y_i + b') \right] y_i = 0$$

$$\sum_{i=1}^{n} \left[x_i - (a'y_i + b') \right] = 0$$

La droite recherchée passe par le point moyen ($\overline{x}, \overline{y}$), qui est donc le point d'intersection des deux droites d'ajustement

$b' = \overline{x} - a'\overline{y}$.

La pente de la droite a' est :

$$a' = \frac{\sum_{i=1}^{n} x_i y_i - n\, \overline{\mathrm{x}}\, \overline{\mathrm{y}}}{\sum_{i=1}^{n} \mathrm{y}_\mathrm{i}^2 - n\, \overline{\mathrm{y}}^2}.$$

Cette expression est le rapport de la covariance de X et Y et de la variance de Y.

$$a' = \frac{Cov(X,Y)}{V(Y)}$$

$$(\hat{x}_i - \overline{x}) = \frac{Cov(X,Y)}{V(X)} (y_i - \overline{y})$$

Dans un tableau de contingence en utilisant les résultats intermédiaires les formules sont :

$$a = \frac{E - \frac{AB}{n}}{C - \frac{A^2}{n}} \ ; \ a' = \frac{E - \frac{AB}{n}}{D - \frac{B^2}{n}} \ ; \ b = \frac{B}{n} - a\frac{A}{n} \ ; \ b' = \frac{A}{n} - a'\frac{B}{n}.$$

Exemple numérique pour des observations individualisées

Soient pour huit années les observations suivantes pour deux grandeurs économiques X et Y, existe-t-il une liaison linéaire entre les valeurs des deux grandeurs.

Tableau 25. Évolution du PIB et de la dépense de consommation finale.

Années	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
PIB	1 840	1 890	1 920	1 970	2010	2 020	1 960	2 000	2 040	2 050	2 050
DCF	1 420	1 450	1 485	1 510	1 550	1 560	1 570	1 600	1 610	1 610	1 620

Source INSEE sur une base 2010 (en milliards d'euros)
PIB : Produit intérieur brut
FBCF : Formation brute de capital fixe
DCF : Dépense de consommation finale

Pour faciliter les calculs, il vaut mieux diviser les données par 1000.

Tableau 26. Équation des droites d'ajustement.

PIB	DCF			
x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1,84	1,42	3,3856	2,0164	2,6128
1,89	1,45	3,5721	2,1025	2,7405
1,92	1,49	3,6864	2,2201	2,8608
1,97	1,51	3,8809	2,2801	2,9747
2,01	1,55	4,0401	2,4025	3,1155
2,02	1,56	4,0804	2,4336	3,1512
1,96	1,57	3,8416	2,4649	3,0772
2	1,6	4	2,56	3,2
2,04	1,61	4,1616	2,5921	3,2844
2,05	1,61	4,2025	2,5921	3,3005
2,05	1,62	4,2025	2,6244	3,321
21,75	16,99	43,0537	26,2887	33,6386

$$a = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} \mathrm{x}_\mathrm{i}^2 - n\,\overline{\mathrm{x}}^2} = \frac{33{,}6386 - 11 \cdot 1{,}98 \cdot 1{,}54}{43{,}0537 - 11 \cdot 1{,}98^2} \cong 0{,}93$$

$$b = \overline{y} - a\overline{x} = 1{,}54 - 0{,}93 \cdot 1{,}98 = -0{,}30$$

$\hat{y} = 0{,}93x - 0{,}30$

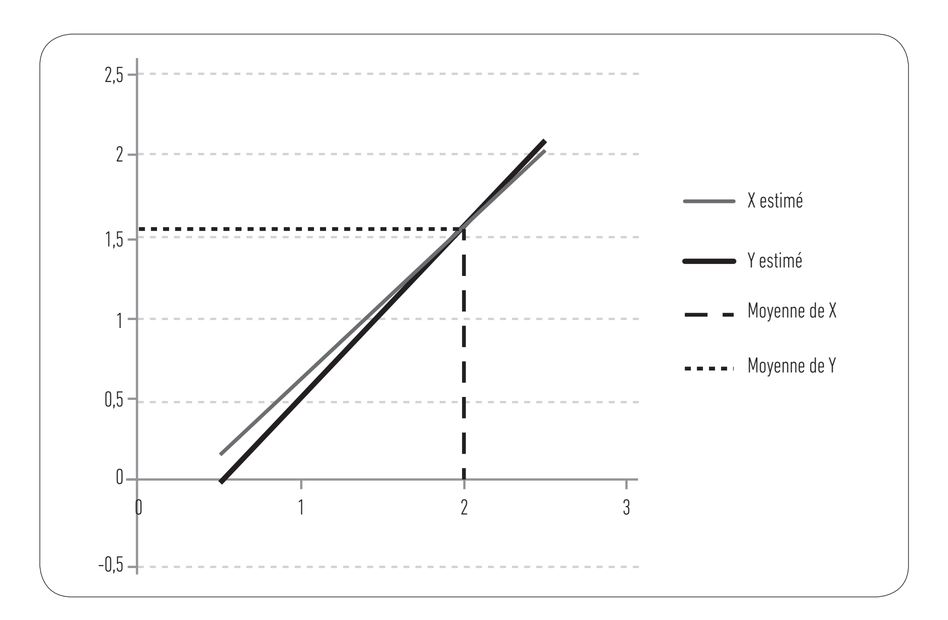
$$a' = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} y_i^2 - n\,\overline{y}^2} = \frac{33{,}6386 - 11 \cdot 1{,}977 \cdot 1{,}54}{26{,}288 - 11 \cdot 1{,}54^2} \cong 0{,}95$$

$b' = \overline{x} - a'\overline{y} = 1{,}98 - 0{,}95 \cdot 1{,}54 = 0{,}50$

$\hat{x} = 0{,}95y + 0{,}5$

Les droites d'ajustement de Y en X et de X en Y passent par le point moyen $(\overline{x}, \overline{y})$ de la distribution.

Figure 9. Représentation graphique des droites d'ajustement.



À l'aide des équations, il est possible avec une hypothèse sur une variable de prévoir la valeur de l'autre.
Avec l'hypothèse du PIB de 22,5, pour calculer la DCF estimée, c'est l'équation $\hat{y} = 0{,}93x - 0{,}30$ qui est pertinente. En effet, x est connu, il reste à déterminer y.

$\hat{y} = 0{,}93x - 0{,}30 = 0{,}93 \cdot 22{,}5 - 0{,}30 = 20{,}625$

Avec l'hypothèse de la DCF de 20, l'équation a utilisée est $\hat{x} = 0{,}95y + 0{,}5$, puisque y est connu.

$\hat{x} = 0{,}95y + 0{,}5 = 0{,}95 \cdot 20 + 0{,}5 = 19{,}5$

L'utilisation de l'équation $\hat{y} = 0{,}93x - 0{,}30$ conduit à une double erreur :

– C'est tout d'abord une erreur logique car x n'est pas connu et la variable expliquée ne peut se substituer à la variable explicative
– Et c'est aussi une erreur de calcul en effet $\hat{y} = 0{,}93x - 0{,}30$ devient $\hat{x} = \frac{y}{0{,}93} + \frac{0{,}3}{0{,}93} = \frac{20}{0{,}93} + \frac{0{,}3}{0{,}93} \cong 21{,}8$.

Le cas des observations groupées en classe

Nous disposons d'informations sur les classes de dépenses de consommation culturelle et de loisirs selon les classes de revenus, les calculs sont réalisés sur les centres de classes. Dans ce cas, les résultats obtenus constituent des évaluations moins précises que dans le cas de données individuelles. Il faut préciser que seuls les ordres de grandeur ont une signification.

Tableau 27. Équation des droites d'ajustement d'une série classée.

Revenus (milliers d'euros)	c_i	**Dépenses loisirs (milliers d'euros)**	$c_i^{'}$	c_i^2	$c_i^{'2}$	$c_i c_i^{'}$
50 à 60	55	3,2 à 4,2	3,7	3 025	13,69	203,5
60 à 62	61	4,2 à 5	4,6	3 721	21,16	280,6
62 à 64	63	5,0 à 5,4	5,2	3969	27,04	327,6
64 à 68	66	5,4 à 5,8	5,6	4 356	31,36	369,6
68 à 73	72	5,8 à 7,0	6,4	5 184	40,96	460,8
73 à 102	91	7,0 à 7,4	7,2	8 281	51,84	655,2
	408		32,7	28 536	186,05	2 297,3

Les valeurs des moyennes :

$\overline{c} = 68$

$\overline{c}' = 5{,}45$

$$a = \frac{\sum_{i=1}^{n} c_i c_i^{'} - n\,\overline{c}\,\overline{c}'}{\sum_{i=1}^{n} c_i^2 - n\,\overline{c}^2} = \frac{2297{,}3 - 6 \cdot 68 \cdot 5{,}45}{28536 - 6 \cdot 68^2} = 0{,}093$$

$b = 5{,}450 \text{ - } 0{,}093 \cdot 68 = \text{-}0{,}88$

L'équation de la droite d'ajustement de c en c'est :

$\hat{c} = 0{,}09c - 0{,}88$

$$a' = \frac{\sum_{i=1}^{n} c_i c_i^{'} - n\,\overline{c}\,\overline{c}}{\sum_{i=1}^{n} c'^2_i - n\,\overline{c}'^2} = \frac{2297{,}3 - 6 \cdot 68 \cdot 5{,}45}{186{,}05 - 6 \cdot 54{,}5^2} = 9{,}4$$

$$b' = 68 - 9{,}4 \cdot 5{,}450 = 16{,}73$$

L'équation de la droite d'ajustement de c en c'est : $\hat{c} = 9{,}4c' + 16{,}73$.

– L'ajustement dans un tableau de contingence

Pour ce faire, il est nécessaire que les deux variables soient quantitatives. Les observations sont rassemblées au centre de classes et les valeurs des variables statistiques sont les centres de classes.

La recherche de l'équation de la droite d'ajustement est identique au cas précédent. Nous déterminons les coefficients a et b de la droite d'ajustement de Y en X. Il s'agit de minimiser la somme des carrés des écarts à la droite. Les équations normales sont analogues aux équations précédentes, à une réserve près, c'est l'introduction des fréquences. Les résultats sont formellement les mêmes.

La pente de la droite d'ajustement de Y en X sera : $a = \frac{Cov(x,y)}{V(y)}$.

Une présentation formelle

Nous pouvons utiliser le tableau du chapitre précédent pour calculer les grandeurs qui nous intéressent.

Tableau 28. Tableau des calculs.

	y_i	Total	B_i	D_i	E_i	$\overline{y}_i$	$V_i(y)$
x_i	n_{ij}	$n_{i.}$	$B_i = \sum_{j=1}^{l} n_{ij} y_j$	$E_j = y_j A_j$	$E_i = x_i B_i$	$\overline{y}_i = \frac{B_i}{n_i}$	$V_i(y) = \frac{D_i}{n_i} - \frac{B_i^2}{n_i^2}$
$n_{.j}$	n_{ij}	n	B	D	E	$\overline{y} = \frac{B}{n}$	$V(y) = \frac{D}{n} - \frac{B^2}{n^2}$
A_j	$A_j = \sum_{i=1}^{k} n_{ij} x_i$	A					
C_j	$C_j = \sum_{i=1}^{k} n_{ij} x_i^2$	C					
E_j	$E_j = y_j A_j$	E					
$\overline{x}_j$	$\overline{x}_j = \frac{A_j}{n_j}$	$\overline{x}_j = \frac{A}{n}$					
$V_j(x)$	$V_j(x) = \frac{C_j}{n_j} - \frac{A_j^2}{n_j^2}$	$V(x) = \frac{C}{n} - \frac{A^2}{n^2}$					

Nous obtenons immédiatement la valeur de a et de a' :

$$a = \frac{E - \frac{AB}{n}}{C - \frac{A^2}{n}} \quad b = \frac{B}{n} - a\frac{A}{n}$$

$$a' = \frac{E - \frac{AB}{n}}{D - \frac{B^2}{n}} \quad b' = \frac{A}{n} - a'\frac{B}{n}$$

Un exemple numérique

Tableau 29. Tableau des calculs.

	2	4	5	6	$n_{i.}$	B_i	D_i	E_i	$\overline{y}_i$	$V_i(y)$
3	10	5		0	15	40	120	120	2,667	0,889
5	15	20	10	5	50	190	810	950	3,800	1,760
7	0	0	40	15	55	290	1 540	2 030	5,273	0,198
$n_{.j}$	25	25	50	20	120	520	2 470	3 100	4,333	1,806
A_j	105	115	330	130	680					
C_j	465	545	2 210	860	4 080					
E_j	210	460	1 650	780	3 100					
$\overline{x}_j$	4,20	4,60	6,60	6,50	5,67					
$V_j(x)$	0,96	0,64	0,64	0,75	1,89					

$$a = \frac{E - \frac{AB}{n}}{C - \frac{A^2}{n}} = \frac{3100 - \frac{680 \cdot 520}{120}}{4080 - \frac{680^2}{120}} \cong 0,68$$

$$b = \frac{B}{n} - a\frac{A}{n} = \frac{520}{120} - 0,676\frac{680}{120} \cong 0,5$$

$$\hat{y} = 0,676x + 0,5$$

$$a' = \frac{E - \frac{AB}{n}}{D - \frac{B^2}{n}} = \frac{3100 - \frac{680 \cdot 520}{120}}{2470 - \frac{520^2}{120}} \cong 0,71$$

$$b' = \frac{A}{n} - a'\frac{B}{n} = \frac{680}{120} - 0{,}708 \cdot \frac{520}{120} = 2{,}6$$

$$\hat{x} = 0{,}708y + 2{,}6$$

La corrélation linéaire

Si les valeurs ou la moyenne d'une variable Y dépendent statistiquement des valeurs d'une variable X, alors les deux variables sont en corrélation. L'ajustement linéaire donne une estimation de la relation entre les deux variables, autrement dit de l'effet de la variation d'une grandeur sur l'autre. Le coefficient de corrélation linéaire estime la qualité de l'ajustement linéaire obtenu par la méthode des moindres carrés. Cette qualité est mesurée par le coefficient de détermination linéaire de Pearson. Il mesure l'intensité de la dépendance des deux variables, il indique aussi dans quelle mesure les variations d'une variable expliquent celles de l'autre. Cette mesure s'appuie sur une analyse de la variance de la variable de Y qui est en partie expliquée par les valeurs prises par la variable X, suivant la relation formalisée par l'ajustement linéaire entre les deux variables. En pratique, il apparaît souvent que le calcul de la corrélation précédera celui des droites d'ajustement. En effet, dans le cas d'un niveau trop faible de validité de la liaison, un calcul d'ajustement n'aurait qu'une signification réduite.
Le coefficient de détermination linéaire retient la part de la variance de Y expliquée par l'ajustement sur la variance totale de Y.

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

En ajoutant et en retirant $\hat{y}_i$ avec $\hat{y}_i = ax_i + b$, il est possible de faire apparaître les différences entre la variable et son estimation, la formule se modifie comme suit :

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})\right]^2 \;;$$

- $V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$;
- $\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = -\overline{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) + \sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i = 0$.

Par application des équations de Gauss, cela donne :

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \ ; \ \hat{y}_i = ax_i + b \ ; \ \sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = 0$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)(ax_i + b) = a\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i + b\sum_{i=1}^{n}(y_i - \hat{y}_i).$$

La variance de Y est la somme de deux termes :

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2.$$

Le premier terme représente la dispersion des valeurs autour de la droite d'ajustement ; c'est la variance non expliquée par l'ajustement ou variance résiduelle. Le second terme exprime la dispersion des valeurs estimées autour de la moyenne, c'est la variance expliquée par l'ajustement.

Le coefficient de détermination s'obtient facilement par $r^2 = \frac{\text{variance expliquée}}{\text{variance totale}}$.

Cela donne, de manière formelle, cette formule :

$$r^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Calculons la valeur du rapport de détermination linéaire :

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{V(y)}$$

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + \overline{y} - a\overline{x} - \overline{y})^2$$

$$\frac{1}{n}\sum_{i=1}^{n}(ax_i + \overline{y} - a\overline{x} - \overline{y})^2 = \frac{a^2}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2V(x)$$

donc :

$$r^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{V(y)} = \frac{a^2V(x)}{V(y)} = \frac{Cov(x,y)^2}{V(x)^2} \times \frac{V(x)}{V(y)} = \frac{Cov(x,y)^2}{V(x)V(y)}$$

$$r^2 = \frac{Cov(x,y)^2}{V(x)\cdot V(y)}$$

$$r^2 = \frac{\left[\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right]^2}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \overline{y})^2\right]}$$

$$r^2 = \frac{\left[\sum_{i=1}^{n} x_i^2 - n\overline{xy}\right]^2}{\left[\sum_{i=1}^{n} x_i^2 - k\overline{x}^2\right]\left[\sum_{i=1}^{k} y_i^2 - n\overline{y}^2\right]}$$

Le coefficient de détermination linéaire est évidemment compris entre 0 et 1 et la liaison sera d'autant plus forte que le coefficient de détermination sera proche de 1.

Ce coefficient signifie que les variations de la variable Y dépendent, pour partie, des variations de la variable X. Les variables X et Y jouant un rôle symétrique, ce sont peut-être les variations de X qui sont expliquées par les variations de Y ; un coefficient de détermination non nul signifie que les variations de X et de Y sont liées. En raison de sa symétrie, le coefficient r^2 mesure l'intensité de la liaison de Y en X tout autant que la liaison de X en Y. Le coefficient de détermination est sans dimension, il représente la fraction des fluctuations de la variable dépendante (Y respectivement. X) qui s'expliquent par les modifications de la variable indépendante (X respectivement Y). Plus le coefficient r^2 est proche de 1, plus la part de la variance expliquée est forte. L'importance de la relation est fournie par le coefficient de détermination qui est interprété comme le pourcentage de la variance expliquée par l'ajustement. Un coefficient de détermination de 0,64 traduit le fait que 64 % de la variation de Y est expliquée par X donc que 36 % de la dispersion reste inexpliquée par l'ajustement. L'ajustement est, souvent, considéré significatif, si r2 est supérieur ou égal à 0,5, où que r le coefficient de corrélation linéaire est supérieur ou égal à 0,7.

Le coefficient de corrélation linéaire ou coefficient de Bravais-Pearson

Dans la pratique, le calcul du coefficient de corrélation linéaire, noté r, précède fréquemment celui du coefficient de détermination linéaire. En effet, il est plus facile à calculer que le coefficient de détermination et par son signe donne le sens de la liaison. Un coefficient positif indique que l'évolution des deux variables se fait dans le même sens, un coefficient négatif que les variables évoluent en sens opposés. Le coefficient de détermination est alors simplement le carré du coefficient de corrélation.

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y} \; ; \; r = \frac{\sum_{i=1}^{i=n}(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x \sigma_y}$$

En pratique, pour obtenir le coefficient de corrélation linéaire, on emploie la formule suivante :

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}.$$

Il est parfois plus aisé d'utiliser les sommes

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} = \frac{\sum_{i=1}^{n} x_i y_i - n\frac{\sum_{i=1}^{n} x_i}{n}\frac{\sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2}}$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

Dans le cas d'un tableau de contingence, la formule devient :

$$r = \frac{E - \frac{AB}{n}}{\sqrt{C - \frac{A^2}{n}} \cdot \sqrt{D - \frac{B^2}{n}}}$$

Le coefficient de détermination peut être calculé à l'aide des coefficients des droites de régression puisque :

$$a = \frac{Cov(x,y)}{V(x)} \;,\; a' = \frac{Cov(x,y)}{V(y)} \;;$$

$$a \cdot a' = \frac{Cov(x,y)}{V(x)} \cdot \frac{Cov(x,y)}{V(y)} = r^2 .$$

Le coefficient de détermination est sans dimension, il représente la fraction des fluctuations de la variable dépendante Y qui s'explique par les modifications de la variable indépendante X.

L'analyse de la corrélation n'établit aucune relation de causalité. Le coefficient de détermination explicite la proportion de la variation qui pourrait être

expliquée si une relation entre les deux variables existait. Les fluctuations de la variable dépendante sont expliquées, non causées par les mouvements de la variable indépendante.
Si le modèle linéaire est satisfaisant en première approche, pour des analyses plus complexes, les économètres préfèrent utiliser des modèles logistiques ou exponentiels.
Nous reprendrons l'exemple de l'évolution du PIB et de la dépense de consommation finale (DCF).

Tableau 30. Évolution du PIB et de la dépense de consommation finale (DCF).

Années	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
PIB	1 840	1 890	1 920	1 970	2 010	2 020	1 960	2 000	2 040	2 050	2 050
DCF	1 420	1 450	1 485	1 510	1 550	1 560	1 570	1 600	1 610	1 610	1 620

Source Insee base 2010 en milliards d'euros
PIB : Produit intérieur brut
FBCF : Formation brute de capital fixe
DCF : Dépense de consommation finale
Pour faciliter les calculs, il vaut mieux diviser les données par 1000.

Tableau 31. Tableau des calculs.

Pour faciliter les calculs, les données ont été divisées par 1 000.

	PIB	DCF			
Année	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
2003	1,84	1,42	3,3856	2,0164	2,6128
2004	1,89	1,45	3,5721	2,1025	2,7405
2005	1,92	1,49	3,6864	2,2201	2,8608
2006	1,97	1,51	3,8809	2,2801	2,9747
2007	2,01	1,55	4,0401	2,4025	3,1155
2008	2,02	1,56	4,0804	2,4336	3,1512
2009	1,96	1,57	3,8416	2,4649	3,0772
2010	2	1,6	4	2,56	3,2
2011	2,04	1,61	4,1616	2,5921	3,2844
2012	2,05	1,61	4,2025	2,5921	3,3005
2013	2,05	1,62	4,2025	2,6244	3,321
Total	21,75	16,99	43,0537	26,2887	33,6386

Le coefficient de corrélation linéaire est facile à calculer

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{33{,}6386 - \frac{21{,}75 \cdot 16{,}99}{11}}{\sqrt{43{,}0537 - \frac{21{,}75^2}{11}} \sqrt{26{,}2887 - \frac{16{,}99^2}{11}}} \cong 0{,}942969644$$

Le coefficient de détermination linéaire donne le niveau de signification de la liaison entre les deux agrégats, c'est le carré du coefficient de corrélation.

$r^2 = 0{,}88919175$

L'ajustement linéaire explique 89 % de la variance des agrégats. L'évolution de la dépense de consommation est expliquée pour 89 % par l'évolution du PIB. L'évolution d'un agrégat s'explique par l'évolution de l'autre sans que l'analyse statistique puisse indiquer le sens de l'influence. En raison de l'importance du coefficient de détermination, il est légitime de calculer les équations d'ajustement entre les deux valeurs.

Tableau 32. Coefficient de détermination.

Revenus (milliers d'euros)	c_i	**Dépenses de loisirs (milliers d'euros)**	c'_i	c_i^2	c'^2_i	$c_i c'_i$
50 à 60	55	3,2 à 4,2	3,7	3 025	13,69	203,5
60 à 62	61	4,2 à 5	4,6	3 721	21,16	280,6
62 à 64	63	5,0 à 5,4	5,2	3 969	27,04	327,6
64 à 68	66	5,4 à 5,8	5,6	4 356	31,36	369,6
68 à 73	72	5,8 à 7,0	6,4	5 184	40,96	460,8
73 à 102	91	7,0 à 7,4	7,2	8 281	51,84	655,2
	408		32,7	28 536	186,05	2 297,3

$$r = \frac{\sum_{i=1}^{n} c_i c'_i - \frac{1}{n} \sum_{i=1}^{n} c_i \sum_{i=1}^{n} c'_i}{\sqrt{\sum_{i=1}^{n} c_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} c_i\right)^2} \sqrt{\sum_{i=1}^{n} c'^2_i - \frac{1}{n}\left(\sum_{i=1}^{n} c'_i\right)^2}}$$

$$r = \frac{2297{,}3 - \frac{1}{6}408 \cdot 32{,}7}{\sqrt{28536 - \frac{1}{6}(408)^2}\sqrt{186{,}05 - \frac{1}{6}(327{,}7)^2}} = 0{,}93558962 \cong 0{,}94$$

$$r^2 = 0{,}875327944 \cong 0{,}87$$

Selon les données analysées, les revenus expliquent 87 % des dépenses de loisirs.

L'ajustement dans un tableau de contingence

Il est nécessaire que les deux variables soient quantitatives. Les observations sont rassemblées au centre de classes, les valeurs des variables statistiques sont les centres de classes.

La recherche de l'équation de la droite d'ajustement est identique au cas précédent. Nous déterminons les coefficients a et b de la droite d'ajustement de Y en X. Il s'agit de minimiser la somme des carrés des écarts à la droite. Les équations normales sont analogues aux équations précédentes, à une réserve près, c'est l'introduction des fréquences. Les résultats sont formellement les mêmes.

La pente de la droite d'ajustement de Y en X sera : $a = \frac{Cov(X,Y)}{V(X)}$.

Une présentation formelle

Nous pouvons utiliser le tableau du chapitre précédent pour calculer les grandeurs qui nous intéressent.

Tableau 33. Tableau des calculs.

	y_i	**Total**	B_i	D_i	E_i	$\overline{y}_i$	$V_i(y)$
x_i	n_{ij}	$n_{i.}$	$B_i = \sum_{j=1}^{l} n_{ij} y_j$	$E_j = y_j A_j$	$E_i = x_i B_i$	$\overline{y}_i = \frac{B_i}{n_i}$	$V_i(y) = \frac{D_i}{n_i} - \frac{B_i^2}{n_i^2}$
$n_{.j}$	n_{ij}	n	B	D	E	$\overline{y} = \frac{B}{n}$	$V(y) = \frac{D}{n} - \frac{B^2}{n^2}$
A_j	$A_j = \sum_{i=1}^{k} n_{ij} x_i$	A					
C_j	$C_j = \sum_{i=1}^{k} n_{ij} x_i^2$	C					
E_j	$E_j = y_j A_j$	E					
$\overline{x}_j$	$\overline{x}_j = \frac{A_j}{n_j}$	$\overline{x}_j = \frac{A}{n}$					
$V_j(x)$	$V_j(x) = \frac{C_j}{n_j} - \frac{A_j^2}{n_j^2}$	$V(x) = \frac{C}{n} - \frac{A^2}{n^2}$					

$$r = \frac{E - \frac{AB}{n}}{\sqrt{C - \frac{A^2}{n}} \cdot \sqrt{D - \frac{B^2}{n}}}$$

Un exemple numérique

Tableau 34. Tableau des calculs.

	2	4	5	6	$n_{i.}$	B_i	D_i	E_i	$\bar{y}_i$	$V_i(y)$
3	10	5		0	15	40	120	120	2,667	0,889
5	15	20	10	5	50	190	810	950	3,800	1,760
7	0	0	40	15	55	290	1 540	2 030	5,273	0,198
$n_{.j}$	25	25	50	20	120	520	2 470	3 100	4,333	1,806
A_j	105	115	330	130	680					
C_j	465	545	2 210	860	4 080					
E_j	210	460	1 650	780	3 100					
$\bar{x}_j$	4,20	4,60	6,60	6,50	5,67					
$V_j(x)$	0,96	0,64	0,64	0,75	1,89					

$$\hat{y} = 0{,}676x + 0{,}5$$

$$\hat{x} = 0{,}708y + 2{,}6$$

$$r = \frac{E - \frac{AB}{n}}{\sqrt{C - \frac{A^2}{n}} \cdot \sqrt{D - \frac{B^2}{n}}} = \frac{3100 - \frac{680 \cdot 520}{120}}{\sqrt{4080 - \frac{680^2}{120}} \cdot \sqrt{2470 - \frac{520^2}{120}}} = 0{,}69190536 \cong 0{,}69$$

$$r^2 = 0{,}691905363^2 = 0{,}478733032 \cong 0{,}48$$

La relation entre les droites d'ajustement et r

Le coefficient de détermination linéaire s'écrit en fonction des pentes des droites d'ajustement a et a'. La relation est très simple : $r^2 = a \cdot a'$.

En effet $a = \frac{Cov(x,y)}{V(x)}$ et $a' = \frac{Cov(x,y)}{V(y)}$.

En remplaçant a et a' par leur valeur en fonction de r :

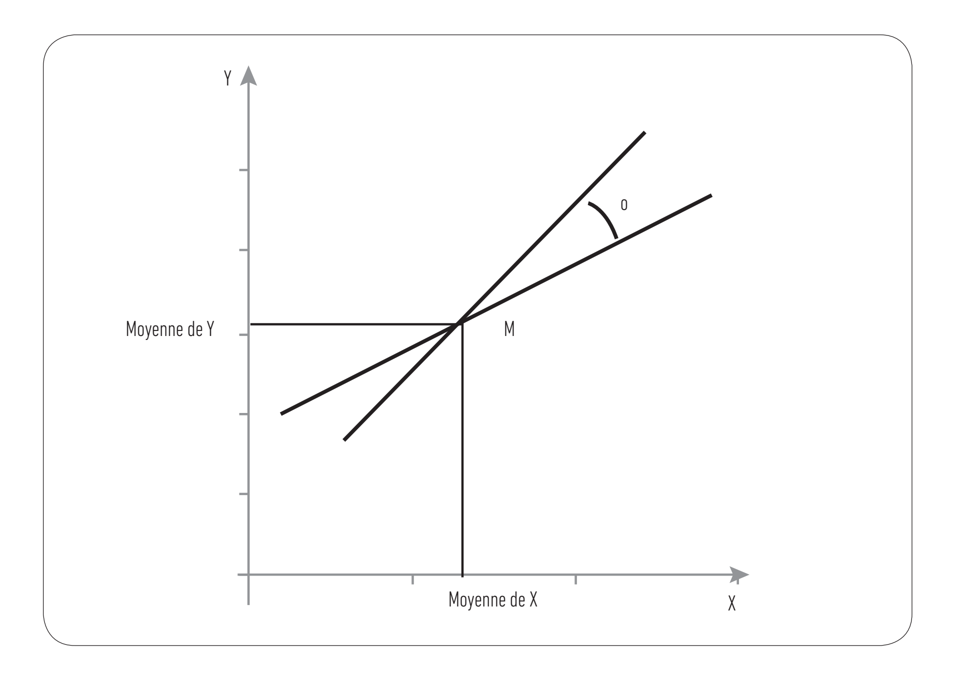
$$a = r\frac{\sigma_y}{\sigma_x} \text{ et } a' = r\frac{\sigma_x}{\sigma_y} \;;$$

$$y - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x}) \;;$$

$$x - \overline{x} = r\frac{\sigma_x}{\sigma_y}(y - \overline{y}).$$

Les pentes des droites d'ajustement sont de même signe : celui de r. De plus, puisque r ≥ 0 alors X et Y varient dans le même sens ; si r ≤ 0, X et Y varient en sens contraire.
La représentation graphique des deux droites d'ajustement montre un angle φ l'angle d'ajustement.

Figure 10. Relation entre les droites d'ajustement et le coefficient de détermination.



Le coefficient de corrélation linéaire est le cosinus de φ r= cos(φ). Plus l'angle d'ajustement est faible plus la corrélation linéaire est forte. Ceci nous explique d'une autre manière pourquoi r est nécessairement compris entre 1 et -1.
Il n'est pas toujours utile de calculer les coefficients a et a' directement, puisque le calcul de r implique la détermination s_x, s_y, ce qui permet d'obtenir immédiatement les coefficients a et a'.

Les outils présentés dans ce chapitre permettent de mettre en lumière les relations entre des variables et de tester la pertinence de celles-ci. Nous avons uniquement traité le cas de deux variables, une extension à un plus grand nombre de variables nécessite du point de vue théorique d'utiliser le calcul matriciel et en pratique de recourir aux moyens de calcul numérique existants. Dans une perspective de formation, ce chapitre est une introduction aux techniques de l'économétrie, domaine largement utilisé dans le champ de l'économie appliquée pour tester et valider - ou invalider - de multiples hypothèses sur les relations entre variables et leur niveau de signification.

Chapitre 4

Les séries chronologiques

Une série statistique qui intègre une dimension temporelle est une chronique. Elle repère la valeur d'une caractéristique d'un phénomène économique à travers le temps. Sa caractéristique essentielle est une dépendance des phénomènes étudiés vis-à-vis du temps. Du point de vue formel, cela se traduit par le fait qu'une chronique est une distribution statistique à deux dimensions dont l'une est le temps. Cependant, le temps joue un rôle particulier en économie et en gestion, dans la mesure où les activités se déroulent dans un temps irréversible. L'ordre des valeurs est fonction du déroulement du temps et en outre chaque valeur dépend plus ou moins des valeurs précédentes. Les grandeurs économiques évoluent en fonction du temps : en gestion le temps constitue une perte de valeur comme dans le cas des amortissements. Tout décideur doit planifier l'évolution des activités de son entreprise ou de l'économie. Dans ce cadre, l'analyse des séries chronologiques représente un outil permettant de fonder les prévisions. Elles ne peuvent fournir plus que des éclairages sur un avenir possible sous l'hypothèse que les dynamiques en œuvre dans le passé perdurent. L'analyse d'une chronique consiste à mettre en évidence et à classer les facteurs qui influent sur la grandeur considérée. Ces mouvements peuvent être de plus ou moins long terme.

En raison de la périodicité des relevés statistiques la valeur des flux dépend du temps. La variable peut être mesurée à un instant donné ou sur un intervalle de temps. Dans le premier cas, on observe un stock à une date donnée ; dans le second, on mesure un flux entre deux dates. À toute mesure de stock, il est le plus souvent possible d'associer une ou plusieurs mesures de flux (les entrées et les sorties) qui expliquent les modifications de la valeur du stock. La périodicité d'une chronique est la durée qui sépare deux informations : jour, semaine, mois, trimestre, année. L'étude d'une série chronologique exige que les relevés successifs soient comparables. En effet, les sources

d'hétérogénéité sont potentiellement nombreuses ; elles proviennent de divers phénomènes : d'une définition insuffisamment précise du phénomène, d'une variation de l'intervalle de temps élémentaire, des variations des prix, des variations de territoire ou de population, de l'extension de l'Union européenne par exemple, de changement de structure d'un phénomène (les produits consommés aujourd'hui diffèrent de ceux consommés il y a 10 ans). Les comparaisons des valeurs prises par une même grandeur au cours du temps doivent tenir compte de ces modifications.
Le traitement statistique des chroniques impose de tenir compte de leurs spécificités. La valeur à un instant donné d'une chronique dépend de la temporalité de plusieurs dynamiques qu'elles soient pluriannuelles, cycliques, saisonnières ou purement accidentelles. Afin de réduire les irrégularités, les méthodes de lissage parfois associées aux techniques d'ajustement fournissent des évaluations pour mettre en évidence les tendances durables qui se dégagent compatibles avec les données disponibles. *A contrario*, les chroniques sont également sensibles à des fluctuations dites saisonnières au cours de l'année ou à des fluctuations dites cycliques sur des périodes de quelques années.

Les composantes d'une série chronologique

Classiquement, les évolutions d'une chronique sont référées à quatre dynamiques ou composantes : la tendance longue (T), les fluctuations cycliques (C), les variations saisonnières (S), les variations aléatoires ou accidentelles (R).
La tendance longue (T dite aussi *trend*) correspond à des évolutions caractéristiques de long terme invariantes sur des périodes décennales, parfois séculaires. Elle traduit l'allure générale du phénomène. Les facteurs responsables de la tendance longue sont structurels : population, technologies, etc.
Les fluctuations cycliques (C) se déroulent sur le moyen terme, c'est-à-dire sur une période supérieure à une année. Les fluctuations conjoncturelles tirent leur nom de leur caractère plus ou moins cyclique. Elles se déroulent sur quelques années, jusqu'à dix ans. Les facteurs explicatifs sont nombreux et aucun modèle, jusqu'à présent, n'est apparu être tout à fait satisfaisant pour expliquer le mouvement cyclique.

Les variations saisonnières (S) sont des mouvements périodiques infra-annuels récurrents. Les facteurs responsables des fluctuations saisonnières sont les conditions climatiques, les coutumes ou les comportements, on pense à des phénomènes sociaux comme les congés annuels pour la production, ou la fin de la scolarité qui induit une hausse de chômage au cours du troisième trimestre. La saisonnalité, au sens large, peut s'appliquer sur des périodes

plus courtes que l'année, à la semaine comme pour les ventes de biens de consommation dans les grandes surfaces, à la journée à l'instar des rythmes de la circulation automobile.

Enfin, les variations aléatoires ou accidentelles (R) surviennent irrégulièrement. Ces mouvements ne suivent aucun modèle, ils sont imprévisibles. Ces variations aléatoires sont parfois attribuables à des événements particuliers : grèves, guerres, élections… Elles constituent aussi ce qui reste non expliqué, un degré d'ignorance. Nous considérerons que les variations aléatoires sont purement dues au hasard de façon à ne pas en tenir compte dans le calcul des composantes alors que même le résidu permet d'apprécier le résultat d'une action volontaire de l'entreprise ou de l'État ; par exemple, on étudie le résidu pour voir si une campagne de publicité a été efficace.

Il est possible de fournir un modèle des séries chronologiques reposant sur l'hypothèse d'une relation entre les composantes. Cette relation peut s'exprimer par une composition additive des mouvements. Elle peut se formuler ainsi :

$$Y = T+C+S+R\ .$$

Une composition multiplicative s'écrit par exemple :

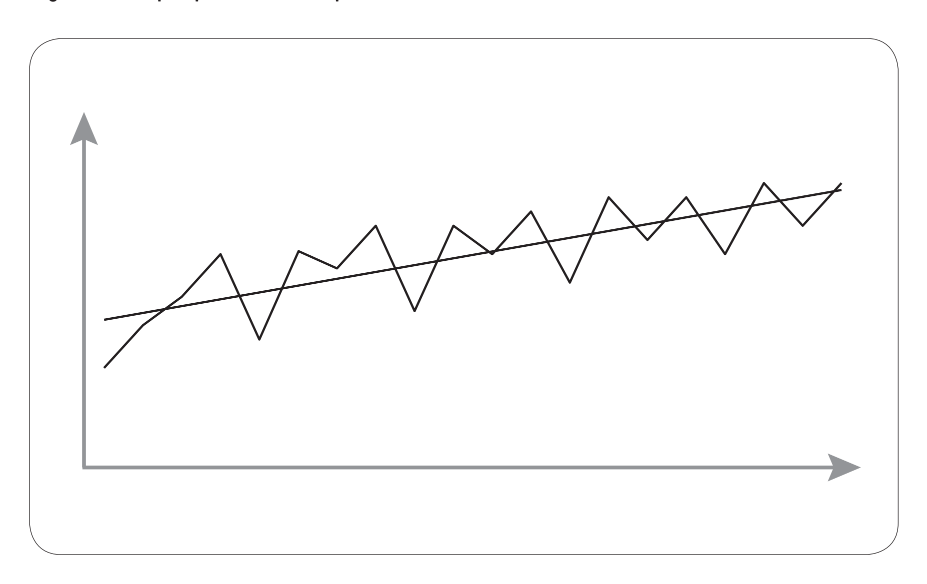
$$Y = T \cdot C \cdot S \cdot R\ .$$

Une composition mixte prendra par exemple cette forme :

$$Y = T \cdot (C + S) + R\ .$$

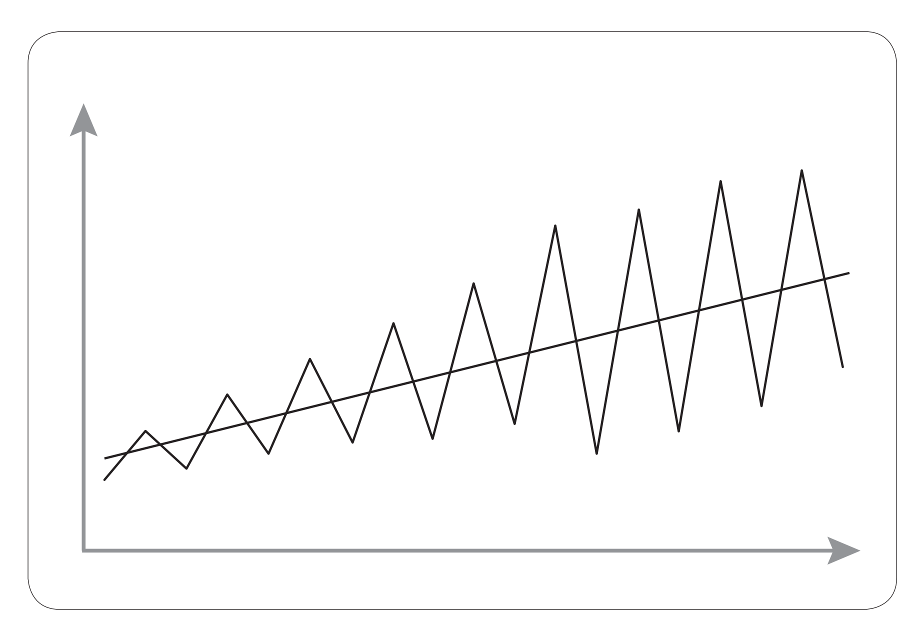
Le choix de la composition des mouvements est souvent délicat. La composition sera additive si la dispersion des fluctuations est peu sensible aux modifications de la variable, les variations étant absolues. Par contre, une composition multiplicative sera retenue si la dispersion augmente avec l'accroissement des valeurs. Les fluctuations sont relatives à l'évolution de la variable. Une représentation graphique permet de percevoir assez facilement le type de composition qui semble le plus pertinent. Le graphique ci-dessous donne un exemple d'une chronique dont la composition des mouvements est additive. En effet, les valeurs fluctuent autour de la droite de régression sans avoir tendance à s'en écarter avec le temps.

Figure 1. Graphique d'une composition additive.



Dans le cas d'une composition multiplicative, les écarts tendent à augmenter avec le temps.

Figure 2. Graphique d'une composition multiplicative.



L'analyse des chroniques privilégie l'étude de la tendance longue et des variations saisonnières.

Les méthodes de lissage

L'objectif central du lissage des chroniques est d'amoindrir l'importance des fluctuations, en particulier accidentelles, de façon à mieux appréhender les évolutions probables en fonction des changements antérieurs. Le lissage peut se substituer quelquefois à la recherche de tendance longue et ce pour deux raisons : soit aucune orientation claire des données ne s'impose soit comme substitut à une recherche de régularité au profit d'une démarche plus empirique. Les techniques de lissage s'utilisent également pour éliminer les mouvements périodiques d'une chronique. Le lissage n'efface le mouvement périodique qu'autant que la période retenue pour le calcul est un multiple de la durée du cycle, et il est nécessaire que le mouvement périodique soit régulier au cours du temps.

La méthode utilisée transforme la distribution initiale en une autre moins irrégulière. Soit une chronique $\{(t_i, y_i)\}$, elle est convertie en une chronique définie par : $\{(t_i, y'_i = f(\cdots, y_{i-1} + y_i + y_{i+1}, \cdots))\}$. Le temps est habituellement le temps médian et la fonction f une moyenne. Les techniques de lissage recourent couramment pour la fonction f à des moyennes arithmétiques faciles à utiliser dans des modèles du fait de leurs définitions algébriques.

Différentes méthodes de regroupement des données sont utilisées pour calculer les moyennes : discontinues ou échelonnées, mobiles, pondérées, d'ordre pair.

Lissage par les moyennes échelonnées

Dans cette méthode, on subdivise l'ensemble des observations en groupes successifs comprenant un même nombre de données. Puis à chaque période au couple (t_i, y_i) se substitue (t_i, y'_i) tel que t soit le temps médian de la période et y' la moyenne arithmétique des y' sur la période considérée.

Tableau 1. Principe des moyennes échelonnées.

Temps	t_1	t_2	t_3	t_4	t_5	t_6
y_i	y_1	y_2	y_3	y_4	y_5	y_6
y'_i		y'_2			y'_5	

Avec : $y'_2 = \frac{y_1 + y_2 + y_3}{3}$ et $y'_5 = \frac{y_4 + y_5 + y_6}{3}$.

Cette méthode soulève trois difficultés :

– 1. Le nombre de chaque période n'est pas fixé. Le choix de 3 est parfaitement conventionnel, le nombre 5 aurait tout aussi bien fait l'affaire ;

– 2. Le choix du premier terme détermine la formation des groupes ;

– 3. Le nombre des données est sensiblement réduit.

Pour des séries très longues, les moyennes échelonnées peuvent être pertinentes précisément du fait de la réduction appréciable du nombre des données. Les moyennes mobiles proposent une solution éliminant la plupart des inconvénients que nous venons d'évoquer.

Lissage par les moyennes mobiles

La méthode de lissage par les moyennes mobiles est une méthode très utilisée. Le principe de calcul est identique à celui des moyennes échelonnées. À une série (t_i, y_i) se substitue une nouvelle série $(t_i, y_i^{'})$ telle que y'_i soit la moyenne arithmétique d'un nombre donné et constant de y_i affectée à t_i le temps médian. Pour une moyenne mobile d'ordre k :

$$y_i^{'} = \frac{y_1 + y_2 + \cdots + y_k}{k} \quad ; \quad y_{i+1}^{'} = \frac{y_2 + \cdots + y_k + y_{k+1}}{k} \; ; \ldots$$

La forme générale est : $M_k(t) = \frac{1}{k}\sum_{i=0}^{k-1} y_{t+i}$.

Tableau 2. Moyennes mobiles sur trois périodes.

t_i	y_i	y'_i
t_1	y_1	
t_2	y_2	$y'_2 = \frac{y_1 + y_2 + y_3}{3}$
t_3	y_3	$y'_3 = \frac{y_2 + y_3 + y_4}{3}$
t_4	y_4	$y'_4 = \frac{y_3 + y_4 + y_5}{3}$
t_i	y_5	

La forme générale est alors

$$y'_t = \frac{1}{3}\left(y_{t-1} + y_t + y_{t+1}\right)$$

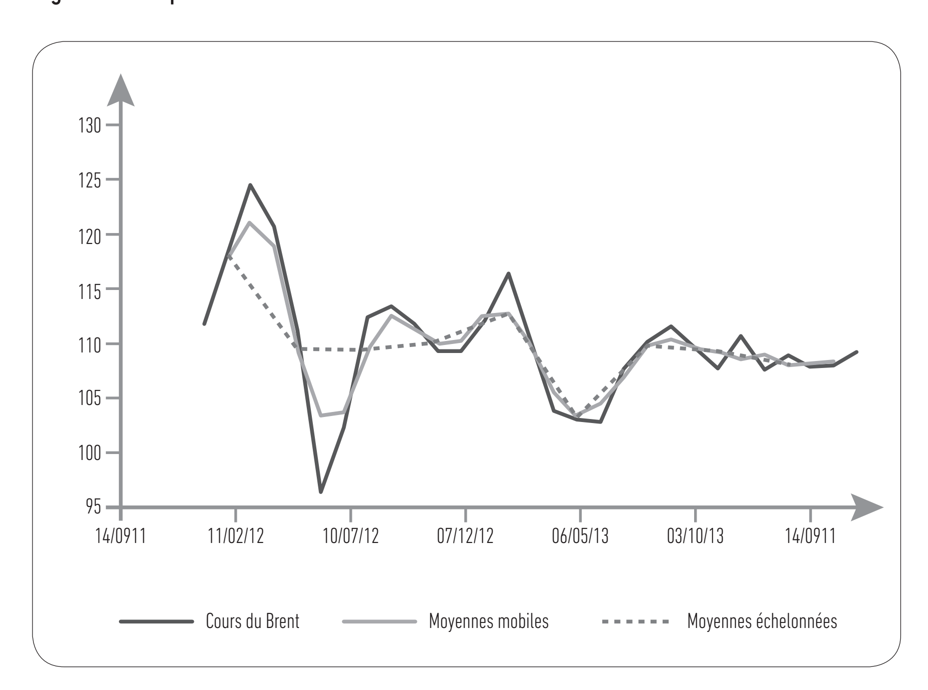
La série des moyennes mobiles comprend moins d'informations que la série originale, deux de moins pour une moyenne mobile d'ordre 3. Cependant, si la série est suffisamment longue, cette perte d'information est mineure. À partir de la distribution des prix du pétrole brut dit Brent, il est possible de calculer des moyennes échelonnées et des moyennes mobiles d'ordre trois.

Tableau 3. Cours en dollar du pétrole Brent.

Mois	Cours	Moyennes échelonnées	Moyennes mobiles
01/01/2012	111,5		
01/02/2012	118,3	118,1	118,1
01/03/2012	124,5		121,2
01/04/2012	120,7		118,9
01/05/2012	111,5	109,5	109,5
01/06/2012	96,4		103,4
01/07/2012	102,3		103,7
01/08/2012	112,4	109,4	109,4
01/09/2012	113,4		112,5
01/10/2012	111,8		111,5
01/11/2012	109,3	110,1	110,1
01/12/2012	109,3		110,2
01/01/2013	112,1		112,6
01/02/2013	116,4	112,7	112,7
01/03/2013	109,7		110,0
01/04/2013	103,8		105,5
01/05/2013	103,0	103,2	103,2
01/06/2013	102,8		104,5
01/07/2013	107,7		106,9
01/08/2013	110,1	109,8	109,8
01/09/2013	111,6		110,4
01/10/2013	109,4		109,6
01/11/2013	107,7	109,3	109,3
01/12/2013	110,7		108,7
01/01/2014	107,6		109,0
01/02/2014	108,7	108,0	108,0
01/03/2014	107,9		108,2
01/04/2014	108,0		108,4
01/05/2014	109,2		

Source : INSEE

Figure 3. Comparaisons des données brutes et lissées.



Le lissage des données est parfaitement visible sur ce graphique. Les prix du Brent sont plus stables en fin de période, les fluctuations se réduisent et les écarts entre le prix et les moyennes se réduisent très nettement.

Les moyennes mobiles pondérées

Les moyennes mobiles pondérées permettent de donner moins d'importance aux valeurs extrêmes. La pondération s'obtient souvent en calculant une seconde série de moyennes mobiles à partir de la première. Les valeurs des pondérations sont généralement les mêmes pour deux valeurs équidistantes d'où la formule :

$$y'_k = a_0 y_k + a_1(y_{k-1} + y_{k+1}) + \cdots\cdots + a_p(y_{k-p} + y_{k+p}) \,.$$

La pondération s'introduit naturellement si à partir d'une première série de moyennes mobiles on en détermine une seconde.

Tableau 4. Moyennes mobiles pondérées sur trois périodes.

t_i	y_i	y_i'	y_i''
t_1	y_1		
t_2	y_2	$y'_2 = \frac{y_1 + y_2 + y_3}{3}$	
t_3	y_3	$y'_3 = \frac{y_2 + y_3 + y_4}{3}$	$y''_3 = \frac{y'_2 + y'_3 + y'_4}{3}$
t_4	y_4	$y'_4 = \frac{y_3 + y_4 + y_5}{3}$	$y''_4 = \frac{y'_3 + y'_4 + y'_5}{3}$
t_5	y_5	$y'_5 = \frac{y_4 + y_5 + y_6}{3}$	
t_6	y_6		

$$y_3'' = \frac{3}{9}y_3 + \frac{2}{9}(y_2 + y_4) + \frac{1}{9}(y_1 + y_5) = \frac{1}{9}(y_1 + 2y_2 + 3y_3 + 2y_4 + y_5)$$

$$y_4'' = \frac{3}{9}y_4 + \frac{2}{9}(y_3 + y_5) + \frac{1}{9}(y_2 + y_6) = \frac{1}{9}(y_2 + 2y_3 + 3y_4 + 2y_5 + y_6)$$

Moyenne mobile d'ordre pair

Ces moyennes sont particulièrement utiles pour supprimer les effets saisonniers des chroniques. Elles seront utilisées spécifiquement pour la recherche des coefficients saisonniers – mensuels ou trimestriels – afin d'annuler l'effet de la saison et de calculer des moyennes sur l'année. Il s'agit d'un exemple de l'utilisation des techniques de lissage d'élimination des fluctuations périodiques pour des chroniques pluriannuelles. Les chroniques à périodicité mensuelle ou trimestrielle nécessitent des calculs de moyennes mobiles sur douze mois ou quatre trimestres.

Au cours de l'année, l'application immédiate des moyennes mobiles sur les données mensuelles ou trimestrielles ne fournit aucun résultat utilisable. En effet, pour une moyenne mobile sur quatre trimestres, pour une valeur qui prend les valeurs y_1, y_2, y_3, y_4. La moyenne est facile à calculer

$$\frac{y_1 + y_2 + y_3 + y_4}{4} = \overline{y}$$

Cependant le temps médian n'existe pas, il se situe entre le deuxième et le troisième trimestre, donc à aucune période de temps de référence.

Pour régler ce problème, les moyennes d'ordre quatre retiennent cinq valeurs. Pour définir un temps médian pertinent, mois ou trimestre, le nombre de valeurs retenues doit être impair tout en assurant un filtre sur l'année.

La formule de la moyenne mobile trimestrielle d'ordre pair retient un nombre impair de trimestres. La moyenne trimestrielle d'ordre pair est calculée sur cinq trimestres, ce qui assure la couverture de l'année ce que ne permettrait pas le choix de trois trimestres. Le résultat du calcul à un trimestre est affecté au trimestre médian. Le calcul de la moyenne mobile mensuelle d'ordre pair utilise les valeurs de treize mois.
La formule générale d'une moyenne mobile d'ordre quatre est la suivante :

$$y'_t = \frac{1}{4}\left(\frac{1}{2}y_{t-2} + y_{t-1} + y_t + y_{t+1} + \frac{1}{2}y_{t+2}\right).$$

Les trimestres sont pondérés par $\frac{1}{4}$ sauf les deux trimestres extrêmes qui le sont par $\frac{1}{8}$.
Le résultat est affecté au temps médian t.
Pour une moyenne mensuelle d'ordre pair, treize trimestres interviennent dans le calcul avec des pondérations de $\frac{1}{12}$ ou de $\frac{1}{24}$.

$$y'_t = \frac{1}{12}(\frac{y_{t+6}}{2} + y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3} + y_{t+4} + y_{t+5} + \frac{y_{t+6}}{2})$$

Ces formules de définition permettent de comprendre la démarche, néanmoins, comme souvent, leur utilisation n'est pas le moyen le plus pratique pour réaliser les calculs surtout en utilisant des tableurs. Il est plus facile et efficace de calculer les moyennes d'ordre pair en passant par des calculs intermédiaires.

Pour une série trimestrielle, la procédure se déroule en deux étapes :

- calcul des sommes mobiles sur quatre trimestres ;
- calcul des moyennes des sommes mobiles.

Première étape : le calcul des sommes mobiles

Tableau 5. Les sommes mobiles sur quatre périodes.

t_i	y_i	S_i
t_1	y_1	$S_1 = y_1 + y_2 + y_3 + y_4$
t_2	y_2	$S_2 = y_2 + y_3 + y_4 + y_5$
t_3	y_3	$S_3 = y_3 + y_4 + y_5 + y_6$
t_4	y_4	$S_4 = y_4 + y_5 + y_6 + y_7$
t_5	y_5	$S_5 = y_5 + y_6 + y_7 + y_8$
t_6	y_6	
t_7	y_7	
t_8	y_8	

Deuxième étape : la moyenne des sommes mobiles :
La moyenne de deux sommes successives est conforme au résultat attendu, c'est-à-dire :

$$y_3' = \frac{1}{8}(S_1 + S_2) = \frac{1}{8}(y_1 + y_2 + y_3 + y_4 + y_2 + y_3 + y_4 + y_5).$$

La division par 8 s'impose en raison des huit valeurs intervenant dans le calcul d'où :

$$y_3' = \frac{1}{4}(\frac{y_1}{2} + y_2 + y_3 + y_4 + \frac{y_5}{2}).$$

Tableau 6. Moyennes mobiles pondérées sur quatre périodes.

t_i	y_i	S_i	y_i'
t_1	y_1	$S_1 = y_1 + y_2 + y_3 + y_4$	
t_2	y_2	$S_2 = y_2 + y_3 + y_4 + y_5$	
t_3	y_3	$S_3 = y_3 + y_4 + y_5 + y_6$	$y_3' = \frac{1}{8}[S_1 + S_2]$
t_4	y_4	$S_4 = y_4 + y_5 + y_6 + y_7$	$y_4' = \frac{1}{8}[S_2 + S_3]$
t_5	y_5	$S_5 = y_5 + y_6 + y_7 + y_8$	$y_5' = \frac{1}{8}[S_3 + S_4]$
t_6	y_6		$y_6' = \frac{1}{8}[S_4 + S_5]$
t_7	y_7		
t_8	y_8		

Moyennes mobiles d'ordre quatre

La série suivante retrace les fluctuations de l'indice trimestriel pour la production industrielle au cours de cinq années.

Tableau 7. Les indices de production industrielle.

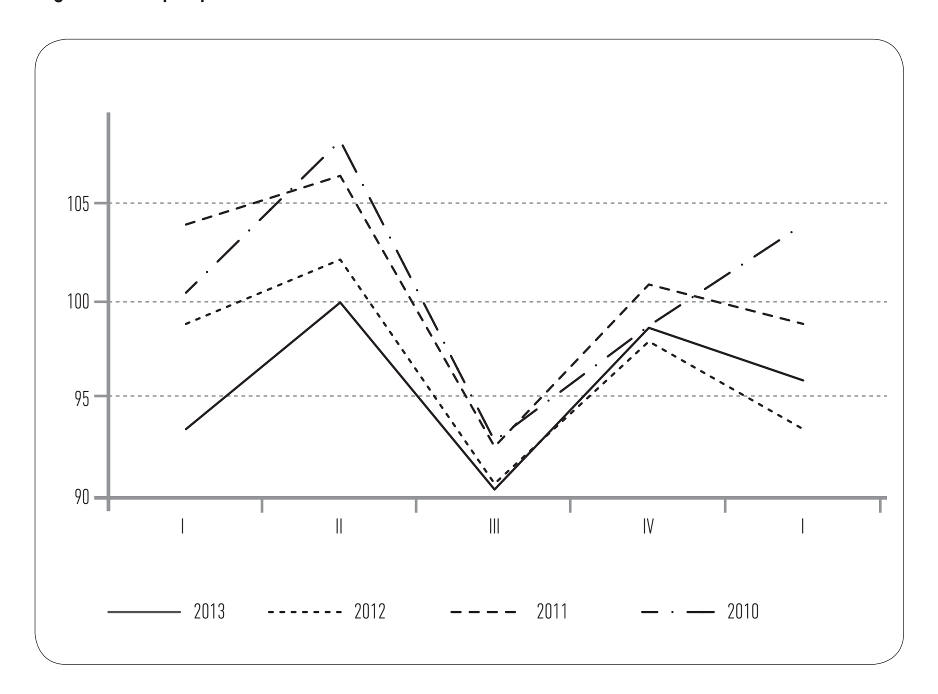
	I	II	III	IV
2014	95,9			
2013	93,4	99,9	90,3	98,6
2012	98,8	102,1	90,6	97,9
2011	103,9	106,4	92,5	100,7
2010	100,4	108,2	92,8	98,7

Source INSEE (calcul des indices trimestriels à partir des indices mensuels)

Une représentation graphique illustre le caractère saisonnier de la série. Le calcul des moyennes mobiles d'ordre quatre de cette série et la représentation sur le même graphique montrent l'effet de lissage de la méthode.

Pour démontrer le caractère saisonnier de cette chronique, une solution est de construire le graphique pour chaque année avec en abscisses les quatre trimestres de chaque année et le premier trimestre de l'année suivante. Le tableau ci-dessus fournit les données nécessaires avec la colonne de gauche qui fournit le niveau de l'indice pour le premier trimestre de l'année suivante.

Figure 4. Graphique des données brutes.



La série montre un profil globalement saisonnier, c'est-à-dire que les variations se produisent, identiques à elles-mêmes, d'une année sur l'autre (périodique de période : un an). La seule « anomalie » est l'augmentation de l'indice entre le quatrième trimestre de l'année 2010 et le premier trimestre de l'année 2011.

Le calcul du *trend* utilise les moyennes mobiles d'ordre 4 afin de lisser les variations sur une année.

$$\begin{cases} S_1 = y_1 + y_2 + y_3 + y_4 \\ S_2 = y_2 + y_3 + y_4 + y_5 \end{cases}$$

$$y'_3 = \frac{1}{8}(S_1 + S_2) = \frac{1}{4}\left(\tfrac{1}{2}y_1 + y_2 + y_3 + y_4 + \tfrac{1}{2}y_5\right)$$

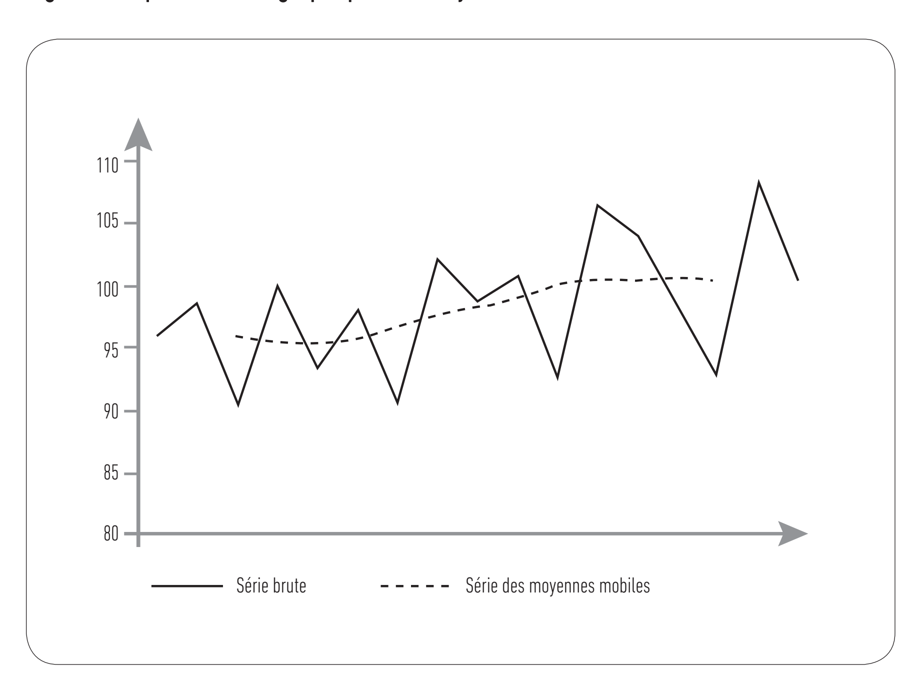
Tableau 8. Tableau des sommes intermédiaires.

	I	II	III	IV
2013	382,2	384,7		
2012	389,4	384	381,8	381,5
2011	403,5	398,4	394,1	392,2
2010	400,1	403,6	401,8	401,5

Tableau 9. Moyennes mobiles.

	I	II	III	IV
2013	95,4	95,5	95,9	
2012	98,3	97,7	96,7	95,7
2011	100,4	100,6	100,2	99,1
2010			100,5	100,7

Figure 5. Représentation graphique des moyennes mobiles.



Le lissage est visible sur ce deuxième graphique, les tendances sont plus nettes, les fluctuations conjoncturelles ou accidentelles ont été en quelque sorte gommées par l'utilisation des moyennes mobiles. Le filtre élimine les variations infra annuelles, le mouvement extra annuel est plus lisible.

La mesure de la saisonnalité

Les mouvements saisonniers sont des mouvements périodiques qui se reproduisent tous les ans à la même époque : par exemple la hausse du chômage au mois d'octobre. La recherche des composantes saisonnières peut avoir deux objectifs, celui d'étudier les mouvements saisonniers ou celui de les éliminer afin de ne garder que les mouvements extra saisonniers. Si un phénomène saisonnier se répète chaque année, il peut cependant évoluer. Il peut être mesuré et dissocié des autres composantes qui influencent le mouvement de la série. Enfin, il peut principalement être causé par des forces exogènes au système économique. Cette dernière caractéristique justifie l'élimination de la saisonnalité dans les statistiques observées. L'élimination des variations saisonnières n'indique pas comment la série aurait évolué en l'absence de saisonnalité, elle révèle le mouvement extra saisonnier. Dans l'étude de la saisonnalité, l'hypothèse la plus simple, que nous retiendrons ici, est de supposer une saisonnalité stable. Des modèles plus complexes retiennent une hypothèse plus appropriée qui suppose des changements de la composante saisonnière.

Quand on procède à une dessaisonalisation, on suppose qu'il n'existe que trois composantes : un mouvement saisonnier $S(t)$, un mouvement extra saisonnier $E(t)$ et des mouvements aléatoires $Z(t)$. Deux modèles sont possibles : le modèle additif et le modèle multiplicatif. On admettra que la fluctuation résiduelle (les mouvements aléatoires) est nulle en moyenne sur l'année et qu'elle n'est pas corrélée avec les autres composantes.

Les deux types de conjonction sont les suivants avec $Y(t)$ pour les données brutes mensuelles :

- forme additive : Y(t) = E(t) + S(t) + Z(t).

Le modèle additif sera retenu si les effets saisonniers semblent s'ajouter à la tendance centrale.

- forme multiplicative : Y(t) = E(t) x S(t) + Z(t).

Dans le cas où les variations saisonnières semblent proportionnelles à la tendance, il est judicieux de retenir le modèle multiplicatif.

Dans tous les cas, l'hypothèse retenue est que le mouvement saisonnier ne joue aucun rôle sur les autres mouvements.

L'étude graphique

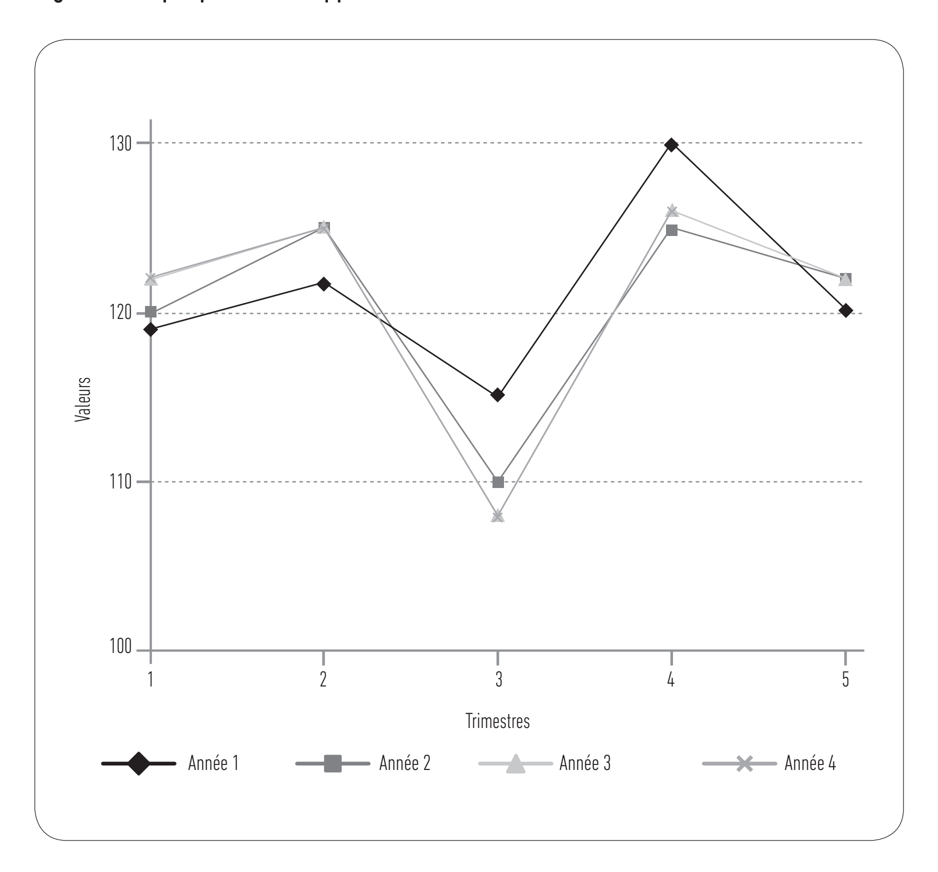
Pour visualiser le mouvement saisonnier, la méthode la plus efficace est de représenter les données sur un graphique annuel. Cette étape permet de

vérifier l'existence et la persistance de mouvements saisonniers. Elle permet aussi d'apprécier si le modèle est additif – l'écart semble constant – ou multiplicatif – l'écart s'accroît.
Le graphique trimestriel montre l'évolution saisonnière de la série. Il est réalisé sur cinq trimestres de façon à faire apparaître l'évolution entre le dernier trimestre de l'année n et le premier trimestre de l'année n-1.

Tableau 11. Indices de production automobile.

Trimestres / Années	I	II	III	IV
1	119	122	115	130
2	120	125	110	125
3	122	125	108	126
4	122	125	108	126

Figure 6. Graphique faisant apparaître la saisonnalité trimestrielle de la distribution.



La dessaisonalisation par la méthode des rapports à la moyenne mobile

Cette méthode est la plus usuelle, car elle ne suppose aucune hypothèse sur les autres mouvements hormis le mouvement aléatoire. La tendance E_{ik} (i l'année, k le mois ou le trimestre) est estimée à l'aide d'une moyenne mobile sur quatre trimestres ou sur douze mois. Deux hypothèses sont nécessaires : que le mouvement saisonnier soit rigoureusement périodique, et que le mouvement aléatoire soit nul en moyenne. La démarche est identique, quel que soit le modèle de composition des mouvements retenus. Pour mener à bien les calculs de dessaisonalisation, on utilise les moyennes mobiles d'ordre pair qui filtrent les mouvements annuels.

Le principe

Les étapes de dessaisonalisation d'une chronique sont les suivantes :

1. vérification graphique du caractère saisonnier de la chronique ;
2. calcul des moyennes mobiles sur 12 mois ou sur 4 trimestres ;
3. calcul des rapports ou des différences aux moyennes mobiles pour obtenir des coefficients pour chaque période ;
4. calcul de la moyenne de ces coefficients pour chaque mois ou trimestre pour obtenir les coefficients bruts ;
5. rectification des coefficients bruts pour obtenir les coefficients normés définitifs ;
6. calcul de la série corrigée des variations saisonnières par division ou soustraction pour chaque mois (respectivement chaque trimestre) de la valeur initiale par le coefficient mensuel (respectivement trimestriel).

Cas d'un modèle additif

Dans le cas du modèle additif, la formule de composition des mouvements prend la forme : $y_{ik} = E_{ik} + S_{ik} + Z_{ik}$ avec E_{ik} le mouvement extra saisonnier, S_{ik} le mouvement saisonnier, Z_{ik} le mouvement aléatoire. L'hypothèse retenue dans cet ouvrage est que la composante aléatoire n'a aucune influence sur les autres composantes. Cette hypothèse est acceptable dans une approche descriptive où les variables sont supposées connues, pour un traitement complet de la question dans un environnement plus aléatoire, elle sera abandonnée. Désormais, la composante Z_{ik} n'apparaîtra plus dans les formules puisque ne jouant aucun rôle dans les calculs.
Comme le mouvement est supposé parfaitement périodique, on peut remplacer les S_{ik} par S_k le coefficient mensuel ou trimestriel donc :
$y_{ik} = E_{ik} + S_k$.

Pour chaque trimestre (respectivement mois), nous calculons les coefficients S'_{ik} de la manière suivante : $y_{ik} - E_{ik} = S'_{ik}$. Nous obtenons ainsi une distribution de coefficients trimestriels (respectivement mensuels). Nous faisons la moyenne arithmétique des coefficients de chaque trimestre (respectivement chaque mois).

$$S'_k = \frac{1}{p}\sum_{i=1}^{p} S'_{ik}$$

Il est facile d'obtenir alors les coefficients saisonniers bruts S'_k. Compte tenu des hypothèses de départ, la somme des coefficients bruts doit être nulle

– pour des coefficients trimestriels : $\sum_{k=1}^{4} S_k = 0$.

– pour des coefficients mensuels : $\sum_{k=1}^{12} S_k = 0$.

Les coefficients saisonniers ayant été calculés indépendamment les uns des autres, ils ne vérifient pas nécessairement cette hypothèse, il faut corriger les coefficients saisonniers. En posant $\bar{S}$ la moyenne des coefficients bruts :

$$\bar{S}' = \frac{1}{n}\sum_{k=1}^{n} S'_k .$$

Les coefficients bruts seront corrigés pour obtenir les coefficients normés définitifs : $S_k = S'_k - \bar{S}'$.

Il est ensuite facile d'obtenir la série(y^*_{ik} ou y^{CVS}_{ik})corrigée des variations saisonnières dite série CVS :

$$y^*_{ik} = y^{CVS}_{ik} = y_{ik} - S_k .$$

Cas d'un modèle multiplicatif

La démarche pour le modèle multiplicatif est semblable. Les hypothèses sur les mouvements aléatoires sont identiques. La forme générale de l'équation est alors : $y_{ik} = E_{ik} \cdot S_{ik}$.

Les coefficients de chaque période se calculent comme le rapport $S'_{ik} = \frac{y_{ik}}{E_{ik}}$.

Nous obtenons une distribution de coefficients, nous calculons la moyenne arithmétique de cette distribution pour chaque mois (respectivement chaque trimestre) pour obtenir les coefficients bruts $S'_k = \frac{1}{p}\sum_{i=1}^{p} S'_{ik}$.

Les coefficients doivent vérifier les hypothèses traduisent dans les équations suivantes (neutralité sur l'année donc la moyenne des coefficients saisonniers est de 1) :

– les S'_k doivent vérifier l'hypothèse : $\sum_{k=1}^{p} S'_k = p$;

– pour des coefficients trimestriels : $\sum_{k=1}^{4} S'_k = 4$;

– pour des coefficients mensuels : $\sum_{k=1}^{12} S'_k = 12$.

– Avec $\overline{S}'$ la moyenne des S'_k: $\frac{1}{p}\sum_{k=1}^{p} S'_k = \overline{S}'$.

– les Sk sont obtenus par rectification : $S_k = \frac{S'_k}{\overline{S}'}$.

La série y^*_{ik} ou y^{cvs}_{ik} corrigée des variations saisonnières série CVS est obtenue facilement avec :

$$y^{CVS}_{ik} = \frac{y_{ik}}{S_k}.$$

Les rapports à la droite de tendance

La méthode est identique à la précédente, le seul changement se situe dans le calcul de la tendance extra saisonnière.
Nous supposons que la fonction extra saisonnière est une fonction affine du temps $E_t = at + b$. La tendance est une droite d'ajustement dont les paramètres sont estimés par la méthode des moindres carrés. Nous obtenons alors l'équation de la droite $E_{ik} = a_{ik} + b$. Ensuite, selon la forme de la composition retenue, il est possible de calculer les coefficients saisonniers bruts ($S'_{ik} = y_{ik} - E_{ik}$ pour une composition additive des mouvements ; $S'_{ik} = \frac{yik}{E_{ik}}$ pour une composition multiplicative).
Il suffit ensuite de procéder comme dans le cas où la tendance extra saisonnière est obtenue par des moyennes mobiles.
Une fois la série désaisonnalisée, il est possible de mettre en lumière la tendance.

Exemple

Pour cet exemple nous reprendrons le cas de la série retraçant les fluctuations de l'indice trimestriel pour la production industrielle au cours de cinq années.

Tableau 12. Les indices de production industrielle.

	I	II	III	IV
2010	100,4	108,2	92,8	98,7
2011	103,9	106,4	92,5	100,7
2012	98,8	102,1	90,6	97,9
2013	93,4	99,9	90,3	98,6
2014	93,4			

Source : INSEE (calcul des indices trimestriels en moyenne d'ordre trois les indices mensuels)

1. Calculez les moyennes mobiles d'ordre quatre de cette série.
2. Représentez graphiquement la nouvelle distribution.

Le caractère saisonnier de cette distribution a été explicité auparavant, il n'y a pas à y revenir.

Solution

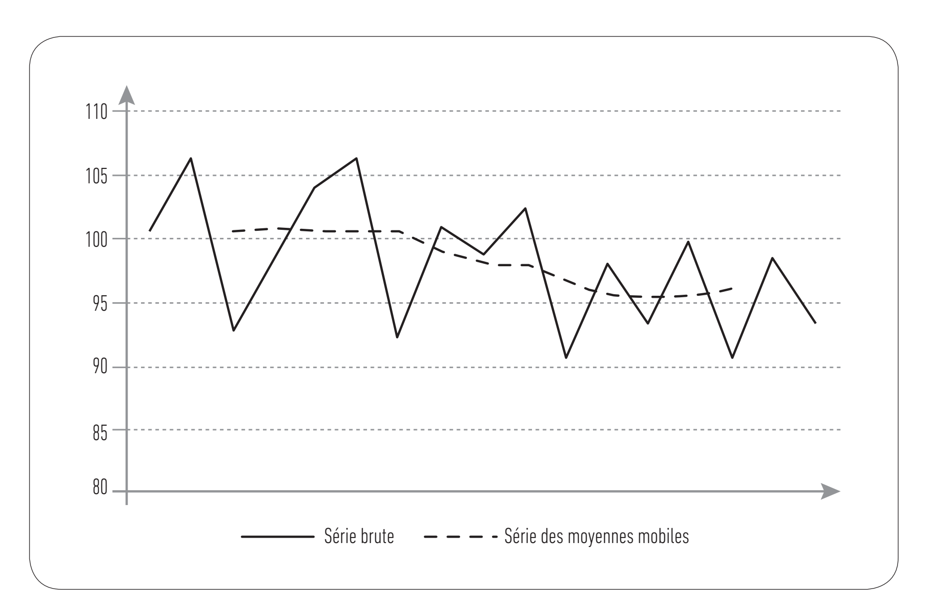
Les moyennes mobiles sont les suivantes, calculées dans l'exemple sur le lissage par les moyennes d'ordre quatre

Tableau 13. Moyennes mobiles.

	I	II	III	IV
2010			100,5	100,7
2011	100,4	100,6	100,2	99,1
2012	98,3	97,7	96,7	95,7
2013	95,4	95,5	95,9	
2014				

La représentation de la série brute peut fournir une indication sur la composition des mouvements de la chronique des indices de production industrielle.

Figure 7. Représentation graphique.



La représentation ne permet pas de choisir sans ambiguïté entre une composition multiplicative ou additive des mouvements. Nous allons donc étudier chacune des compositions que nous avons évoquées.

Hypothèse d'une composition multiplicative

Les coefficients bruts sont obtenus en faisant le rapport entre la valeur brute et la moyenne mobile $S'_{ik} = \frac{y_{ik}}{E_{ik}}$.

Tableau 14. Coefficients bruts (composition multiplicative).

S'_{ik}	I	II	III	IV
2010			92,4	98,0
2011	103,5	105,7	92,3	101,7
2012	100,5	104,5	93,7	102,3
2013	97,9	104,6	94,2	
2014				

Une fois les coefficients bruts obtenus, il faut procéder au calcul des coefficients moyens par trimestre :

$$S'_k = \frac{1}{p}\sum_{i=1}^{p} S'_{ik} \cdot$$

Tableau 15. Moyennes des coefficients bruts par trimestre.

S'_k	I	II	III	IV	$\sum_{k=1}^{4} S'_k$
	100,6	105,0	93,1	100,7	399,4

La somme des coefficients bruts moyens est différente de la somme théorique qui dans le cas d'une distribution trimestrielle est de quatre. Les coefficients doivent être rectifiés en divisant chaque coefficient par la moyenne des coefficients bruts, soit 0,9984. Ce qui permet de calculer les coefficients saisonniers définitifs.

S_k	I	II,	III	IV	$\sum_{k=1}^{4} S_k$
	100,8	105,1	93,3	100,8	400

Il est alors possible de calculer la série corrigée des variations saisonnières $y_{ik}^{CVS} = \frac{y_{ik}}{S_k}$.

Tableau 16. Série CVS composition multiplicative

y^{CVS}	I	II	III	IV
2010	99,6	102,9	99,5	97,9
2011	103,1	101,2	99,2	99,9
2012	98,0	97,1	97,1	97,1
2013	92,7	95,0	96,8	97,8
2014	92,7			

Il est alors possible de calculer la valeur dessaisonalisée de l'indice de pour le second trimestre de 2014. En supposant que la valeur brute est de 107, la valeur dessaisonalisée est

$$y_{2014/II}^{CVS} = \frac{107}{1,05} = 101,9 .$$

Hypothèse d'une composition additive

Les coefficients bruts sont obtenus par soustraction à la tendance extra saisonnière des moyennes mobiles trimestrielles $y_{ik} - E_{ik} = S'_{ik}$.

Tableau 17. Coefficients saisonniers bruts (hypothèse additive).

	I	II	III	IV
2010			-7,7	-2,0
2011	3,5	5,8	-7,7	1,6
2012	0,5	4,4	-6,1	2,2
2013	-2,0	4,4	-5,6	
2014				

Le calcul de la moyenne des coefficients bruts trimestriels permet de vérifier si la condition théorique est vérifiée, c'est-à-dire que la somme des coefficients est nulle :

S'_k	I	II	III	IV	$\sum_{k=1}^{4} S'_k$
	0,7	4,9	-7,2	0,6	-1,0

La condition théorique n'étant pas vérifiée, les coefficients doivent être rectifiés en soustrayant la moyenne des coefficients bruts (-0,25) à chaque coefficient brut.

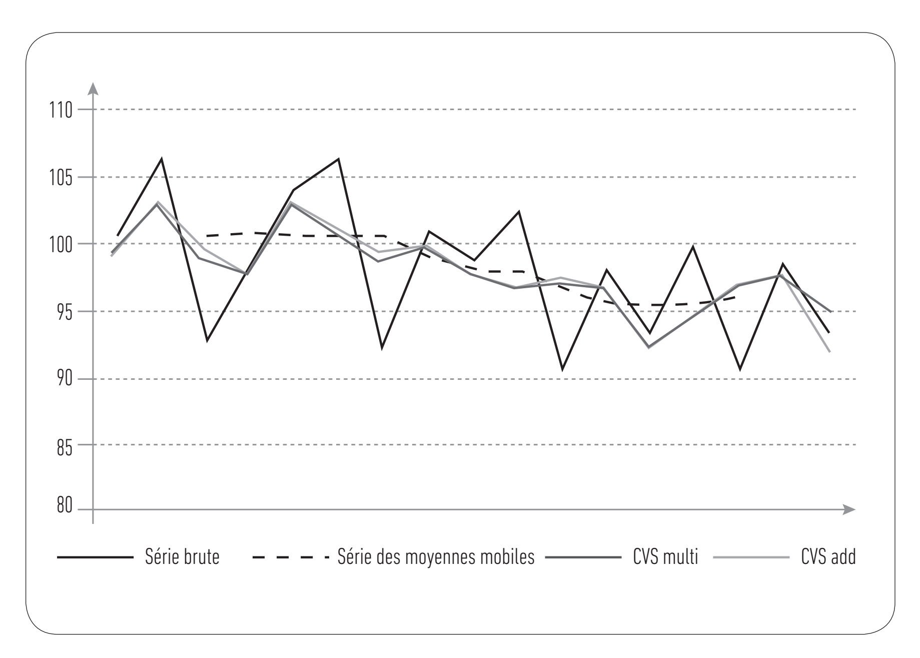
S_k	I	II,	III	IV	$\sum_{k=1}^{4} S_k$
	0,9	5,1	-6,9	0,9	0,0

Tableau 18. Série CVS.

y^{CVS}	I	II	III	IV
2010	99,5	103,1	99,7	97,8
2011	103,0	101,3	99,4	99,8
2012	97,9	97,0	97,5	97,0
2013	92,5	94,8	97,2	97,7
2014	92,5			

Les écarts entre les valeurs des deux séries CVS sont très faibles, les deux séries peuvent être considérées comme identiques.

Figure 8. Graphique des séries.



Exemple d'une correction de variations saisonnières à l'aide d'un trend estimé par une droite d'ajustement

L'objectif est de construire la série CVS de la distribution ci-dessous en estimant le *trend* par une droite d'ajustement.

Tableau 19. Série initiale. y_{ik}

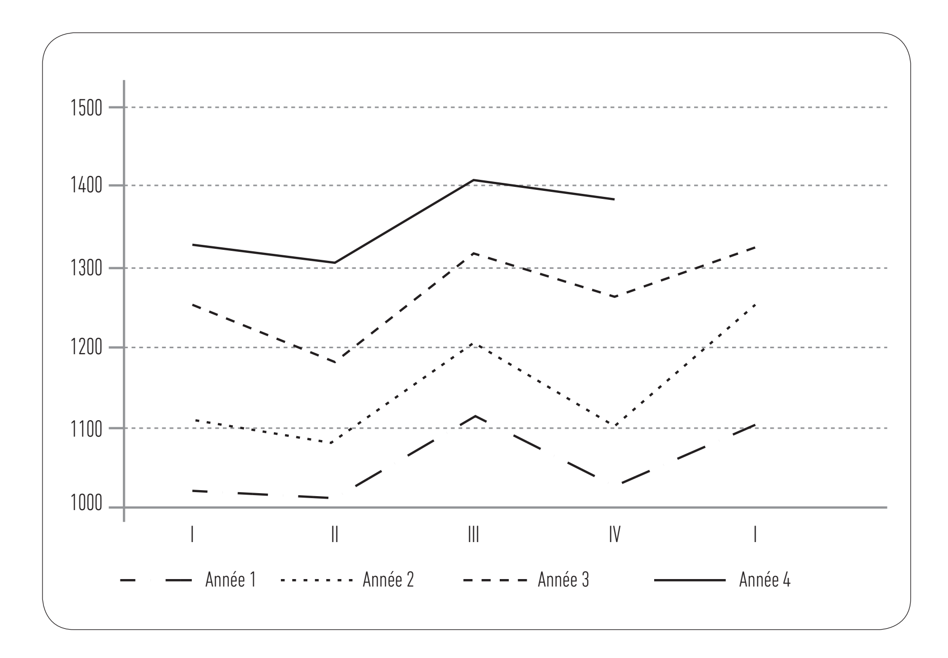
Trimestre / Année	I	II	III	IV
Année 1	1 020	1 010	1 110	1 030
Année 2	1 100	1 080	1 200	1 100
Année 3	1 250	1 180	1 310	1 260
Année 4	1 320	1 300	1 400	1 380

Avant tout calcul, il faut vérifier le caractère saisonnier de la distribution. Une représentation graphique sur cinq trimestres permet de tester cette hypothèse.

Tableau 20. Série pour la représentation graphique.

Année \ Trimestre	I	II	III	IV	I
Année 1	1 020	1 010	1 110	1 030	1 100
Année 2	1 100	1 080	1 200	1 100	1 250
Année 3	1 250	1 180	1 310	1 260	1 320
Année 4	1 320	1 300	1 400	1 380	

Figure 9. Représentation graphique de la distribution.



Le caractère saisonnier est clairement visible sur le graphique.
Pour éviter les effets saisonniers, l'équation de la droite d'ajustement est calculée sur les moyennes annuelles, les temps seront les temps moyens pour chacune des années avec l'hypothèse que le temps initial (t=1) correspond au premier trimestre de l'année 1.

Le tableau suivant explicite les calculs de l'équation de la droite d'ajustement.

Tableau 21. Ajustement par une droite.

	t	y	t^2	y^2	t-y
Année 1	2,5	1 042,5	6	1 086 806	2 606,25
Année 2	6,5	1 120,0	42	1 254 400	7 280
Année 3	10,5	1 250,0	110	1 562 500	13 125
Année 4	14,5	1 350,0	210	1 822 500	19 575
Total	34	4 763	369	5 726 206	42 586

L'ajustement par une droite consiste à rechercher les coefficients a et b tel que :

$$\hat{y} = at + b$$

$$a = \frac{\sum_{i=1}^{n} y_i t_i - n \cdot \overline{y} \cdot \overline{t}}{\sum_{i=1}^{n} t_i^2 - n\overline{t}^2} = \frac{42586 - 4 \cdot \frac{34}{4} \cdot \frac{4763}{4}}{369 - 4 \cdot \left(\frac{34}{4}\right)^2} = 26,3$$

$$b = \overline{y} - a\overline{t} = \frac{4763}{4} - 26,3 \cdot \frac{34}{4} = 967$$

$$\hat{y} = at + b = 26,3t + 967$$

Il est possible de calculer la série des valeurs obtenue par l'ajustement.

Tableau 22. Valeurs estimées par la droite d'ajustement $\hat{y}_{ik}$.

Année \ Trimestre	I	II	III	IV
Année 1	993,3	1 019,6	1 045,9	1 072,2
Année 2	1 098,5	1 124,8	1 151,2	1 177,5
Année 3	1 203,8	1 230,1	1 256,4	1 282,7
Année 4	1 309,0	1 335,3	1 361,7	1 388,0

Avec une hypothèse de composition multiplicative des mouvements, le tableau des coefficients bruts est obtenu ci-dessous. Les coefficients bruts sous l'hypothèse d'une composition multiplicative des mouvements sont obtenus en faisant le rapport de la valeur brute à la valeur estimée $\frac{y_{ik}}{\hat{y}_{ik}}$.

Pour faciliter la lecture et les calculs, les rapports sont multipliés par 100 pour obtenir les coefficients bruts $S'_{ik} = \frac{y_{ik}}{\hat{y}_{ik}} \cdot 100$, d'où la distribution des coefficients bruts.

Tableau 23. Coefficients bruts S'_{ik}.

Année \ Trimestre	I	II	III	IV
Année 1	102,69	99,06	106,13	96,06
Année 2	100,13	96,01	104,24	93,42
Année 3	103,84	95,93	104,27	98,23
4	100,84	97,35	102,82	99,43

Nous calculons ensuite un indicateur de tendance centrale pour chaque distribution trimestrielle des coefficients bruts en prenant la moyenne arithmétique de ceux-ci, $S'_k = \frac{1}{4}\sum_{i=1}^{4} S'_{ik}$.

Tableau 24. Moyenne des coefficients bruts S'_k.

	I	II	III	IV
S'_k	101,88	97,09	104,36	96,78

Les coefficients définitifs doivent respecter une contrainte, leur somme doit être égale à 400. Si la somme des coefficients moyens bruts est différente, les coefficients devront être rectifiés.

Somme des coefficients bruts moyens, $\sum_{i=1}^{4} S'_k = 400{,}11$, les coefficients doivent être rectifiés :

$\text{Coefficient de rectifcation} = \frac{400{,}11}{400}$.

Tableau 25. Coefficients rectifiés S_k.

	I	II	III	IV
S_k	101,85	97,06	104,33	96,76

Pour obtenir la série CVS, il suffit de diviser les valeurs de la série initiale par les coefficients obtenus. Les valeurs initiales du trimestre k sont divisées par le coefficient S_k.

La forme générale valeur CVS est donc : $y_{ik}^{CVS} = \frac{y_{ik}}{S_k}$.

Tableau 26. Série CVS y_{ik}^{CVS} .

Année \ Trimestre	I	II	III	IV
Année 1	1 002	1 041	1 064	1 065
Année 2	1 080	1 113	1 150	1 137
Année 3	1 227	1 216	1 256	1 302
Année 4	1 296	1 339	1 342	1 426

Les résultats permettent de donner des estimations des valeurs futures.
Il est possible de calculer la valeur réelle en connaissant la valeur dessaisonalisée. Soit 1 428 la valeur dessaisonalisée pour le trimestre III de l'année 5, la valeur réelle estimée est alors :

$y_{5/III} = y_{5/III}^{CVS} \cdot S_{III} = 1428 \cdot 1{,}0433 = 1498{,}8324 \cong 1499$.

La valeur pour le trimestre 3 de l'année 6 (t =23) estimée par la droite d'ajustement est de $\hat{y}_{6,III} = 23 \cdot 26{,}3 + 967 \cong 1572$.

La valeur CVS est alors $y_{6,III}^{CVS} = \dfrac{\hat{y}_{6,III}}{S_{III}} = \dfrac{1572}{1{,}0433} \cong 1507$

La valeur du semestre 2 de l'année 8 de 1 702 sera corrigée des variations saisonnières 1 754

$\left(\dfrac{1702}{0{,}9706} \cong 1754\right)$.

Nous avons arrondi les chiffres à l'unité pour ne pas donner une impression, fausse, d'une grande précision.

La détermination de la tendance (T)

La détermination de la tendance longue ou *trend* permet d'atteindre plusieurs objectifs. Tout d'abord, elle permet d'élaborer des modèles descriptifs des tendances antérieures. Ensuite, si les raisons existent d'un maintien de ces orientations, elle peut aussi permettre de faire des prévisions à structure constante. Enfin éliminer le *trend* offre la possibilité d'étudier les autres mouvements, en particulier les fluctuations cycliques et saisonnières. La détermination du *trend* ne peut se faire que si l'on dispose d'une série d'observations assez longue. La détermination se fait généralement à partir des données annuelles qui éliminent la composante saisonnière et les fluctuations aléatoires. La détermination rigoureuse d'une tendance devrait prendre en compte les phénomènes de l'autocorrélation des données – la valeur du chiffre d'une période dépend des valeurs des périodes précédentes. Les considérations dans cet ouvrage se limitent à une première approche qui pourra être complexifiée une fois les lois de probabilités connues.

La détermination graphique consiste à tracer sur le graphique, à main levée, la courbe régulière qui s'adapte le mieux à l'allure globale de la distribution. Cette procédure, très subjective, offre la possibilité de visualiser la forme générale de l'évolution, elle permet une première appréhension des mouvements.

Utilisation d'un ajustement

La méthode par les ajustements consiste à calculer l'équation de la fonction d'ajustement qui paraît la plus adaptée puis à estimer le niveau de signification. Il est souvent pertinent d'éliminer les fluctuations saisonnières et cycliques avant de calculer l'équation de la fonction d'ajustement.

Ajustement par une droite

L'ajustement par une droite consiste à rechercher les coefficients a et b tel que :
$\hat{y}_t = at + b$
Les coefficients a et b sont déterminés comme dans le cas de l'ajustement entre deux variables quelconques. La détermination de l'équation de la seconde droite d'ajustement n'a qu'un sens d'analyse statistique. Le temps ne peut être déterminé par l'évolution d'une grandeur.

Nous rappelons les principaux résultats obtenus dans le cas de l'ajustement linéaire :

$$a = \frac{\sum_{i=1}^{n} y_i t_i - n\overline{y}\overline{t}}{\sum_{i=1}^{n} t^2 - n\overline{t}^2}$$

$$b = \overline{y} - a\overline{t}$$

ou en réalisant le changement de variable :

$$y'_i = y_i - \overline{y}$$

$$t'_i = t_i - \overline{t}$$

$$a = \frac{\sum_{i=1}^{n} y'_i t'_i}{\sum_{i=1}^{n} t'^2_i}$$

Le coefficient de corrélation linéaire permettra alors de donner un niveau de signification à l'ajustement analytique, la qualité de l'estimation de la valeur de la grandeur en fonction du temps, en quelque sorte, et par là d'indiquer l'importance des fluctuations cycliques ou saisonnières.

$$r^2 = \frac{\left(\sum_{i=1}^{n} y'_i t'_i\right)^2}{\left(\sum_{i=1}^{n} t'^2_i\right)\left(\sum_{i=1}^{n} y'_2\right)}$$

Exemple

L'évolution de l'espérance de vie à la naissance pour les femmes est donnée par le tableau suivant :

Tableau 27. Espérance de vie des femmes.

t	1946	1950	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005	2010
e_v	65,2	69,2	71,5	73,6	74,7	75,9	76,9	78,4	79,4	81,0	81,9	82,8	83,9	84,7

Champ : France métropolitaine, territoire au 31 décembre 2010
Source : Insee, statistiques de l'état civil et estimations de population

L'espérance de vie à la naissance est la moyenne à une date donnée de la distribution des âges au moment de la mort.

Pour calculer l'équation de la droite d'ajustement de l'espérance de vie en fonction du temps, le tableau suivant donne toutes les informations nécessaires.

Tableau 28. Tableau des calculs intermédiaires.

Année	Temps	Espérance de vie			
	t	e_v	t^2	e_v^2	$e_v \cdot t$
1946	1	65,2	1	4 251,04	65,2
1950	5	69,2	25	4 788,64	346
1955	10	71,5	100	5 112,25	715
1960	15	73,6	225	5 416,96	1 104
1965	20	74,7	400	5 580,09	1 494
1970	25	75,9	625	5 760,81	1 897,5
1975	30	76,9	900	5 913,61	2 307
1980	35	78,4	1 225	6 146,56	2 744
1985	40	79,4	1 600	6 304,36	3 176
1990	45	81,0	2 025	6 544,81	3 640,5
1995	50	81,9	2 500	6 707,61	4 095
2000	55	82,8	3 025	6 855,84	4 554
2005	60	83,9	3 600	7 022,44	5 028
2010	65	84,7	4 225	7 174,09	5 505,5
Total	456	1 079,0	20 476	83 579,11	36 671,7

L'année 1946 est le temps t = 1.
L'équation de la droite d'ajustement de l'espérance de vie par rapport au temps a pour équation :

$$a = \frac{\sum_{t=1}^{14} e_{v_t} t - 14\overline{e}_v\overline{t}}{\sum_{t=1}^{14} t^2 - 14\overline{t}^2} = \frac{36671{,}7 - 14 \cdot \frac{456}{14} \cdot \frac{1079}{14}}{20476 - 14 \cdot \left(\frac{456}{14}\right)^2} \cong 0{,}272$$

$$b = \overline{e}_v - a\overline{t} = 1078{,}9 - 0{,}272 \cdot \frac{456}{14} \cong 68{,}2$$

L'équation de l'espérance de vie en fonction du temps s'écrit alors :
$\hat{e}_v = 0{,}272t + 68{,}2$.

Le coefficient de corrélation linéaire se calcule comme suit :

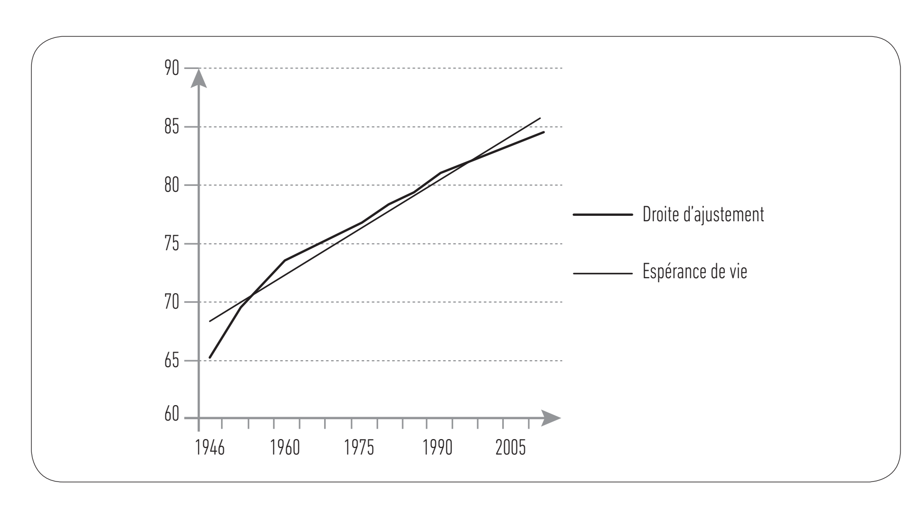
$$r = \frac{\sum_{t=1}^{14} e_{v_t} t - 14\overline{e}_v\overline{t}}{\sqrt{\sum_{t=1}^{14} t^2 - 14\overline{t}^2} \cdot \sqrt{\sum_{t=1}^{14} e_{vt}^{\ 2} - 14\overline{e}_v^{\ 2}}} = \frac{36671{,}7 - 14 \cdot \frac{456}{14} \cdot \frac{1079}{14}}{\sqrt{20476 - 14 \cdot \left(\frac{456}{14}\right)^2} \cdot \sqrt{83579 - 14 \cdot \left(\frac{1079}{14}\right)^2}} = 0{,}9791$$

Le coefficient de corrélation nous indique que l'espérance de vie est croissante au cours du temps.
Le coefficient de détermination linéaire r^2 est d'environ 0,959. Il signifie que l'espérance de vie est « expliquée » statistiquement pour environ 96 % par le temps. Les causes véritables restent à expliciter.

Tableau 29. Ajustement de l'espérance de vie par une droite.

Année	Espérance de vie
	e_v
1946	65,2
1950	69,2
1955	71,5
1960	73,6
1965	74,7
1970	75,9
1975	76,9
1980	78,4
1985	79,4
1990	81,0
1995	81,9
2000	82,8
2005	83,9
2010	84,7

Figure 10. Représentation graphique de la série brute et de la droite d'ajustement.



Dans le cas d'un ajustement polygonal, on recherche les coefficients d'une fonction polygonale générale du type :
$\hat{y}_t = a_0 + a_1 t + a_2 t^2 + \ldots\ldots + a_k t^k$.

Le choix du degré est fonction de la forme générale du nuage de points. Les coefficients sont déterminés par la méthode des moindres carrés.
Lorsque la croissance du phénomène est très prononcée, en hausse ou en baisse, il semble plus pertinent de chercher un ajustement exponentiel. Le taux d'accroissement est $\frac{dy(t)}{dt} = r \cdot y(t)$.

La solution est $y = e^{rt} + c$.

Il est souvent pratique d'exprimer le phénomène en fonction d'un multiplicateur moyen.

– $y_t = y_0(1+r)^t$

En pratique, l'ajustement se ramène à appliquer aux logarithmes les méthodes utilisées pour l'ajustement par une droite.

Principe de l'ajustement par une logistique

Cette courbe fut d'abord utilisée par des démographes avant d'être appliquée aux phénomènes économiques. Les phénomènes ne croissent pas sans contraintes qui sont souvent fonctions de la croissance même du phénomène. La croissance d'une production se heurte aux disponibilités en énergie et en matières premières. La contrainte peut être physique, limitations des quantités, ou économiques, hausse des prix. Elle s'exprime mathématiquement par :

$$\frac{dy(t)}{dt} = r \cdot y(t)\left[1 - \frac{y(t)}{k}\right].$$

La croissance de la grandeur est limitée par sa propre croissance. La courbe représentative a pour expression :

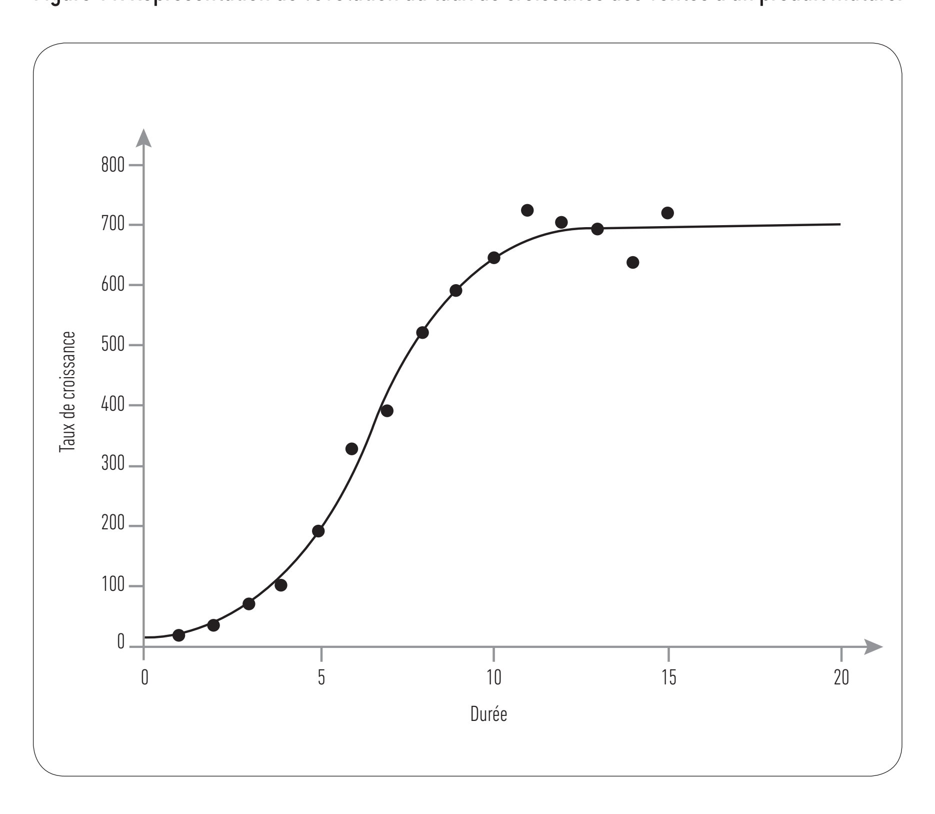
$$y(t) = \frac{k}{1 + ce^{-rt}} .$$

Il est possible de généraliser la tendance logistique avec des équations de la forme :

$$y(t) = \frac{k}{1 + ce^{f(t)}}$$

où f(t) est une fonction quelconque.
La croissance de y(t) est logistique et tend vers une limite k.

Figure 11. Représentation de l'évolution du taux de croissance des ventes d'un produit mature.



Cet ajustement rend de compte de phénomènes à développement rapide au cours d'une première période et qui se freinent ensuite du fait de leur propre croissance.

La composante cyclique

Une fois construite, la série désaisonnalisée révèle parfois d'autres fluctuations autour du mouvement de longue durée. Ces fluctuations portent le nom de cycles. Ces variations sont considérées comme le mouvement économique, alors que la tendance représente l'histoire et que le mouvement saisonnier est conjoncturel. Ces fluctuations entraînent des crises d'où les nombreuses études réalisées pour expliquer le phénomène ou plus modestement pour le décrire, l'objet étant toujours la prévision des dépressions afin d'envisager des politiques contra cycliques. Pour ce faire, il est tout à fait possible d'utiliser les méthodes utilisées pour la dessaisonalisation sur la série désaisonnalisée. Il suffit dans le cas des moyennes mobiles de calculer des moyennes mobiles de la périodicité du cycle mis en lumière ou de procéder aux calculs dans le

cas d'un ajustement par une fonction. Nous présenterons deux méthodes plus spécifiques celle des résidus et celle du cycle moyen. Il est cependant de plus en plus rare que l'on procède à des estimations des coefficients cycliques. L'intervention globale et massive des pouvoirs publics (jamais moins de 40 % de dépenses publiques dans les PIB des pays développés) a réduit l'importance des cycles à la fois dans les réalités économiques que dans l'analyse au profit des études conjoncturelles. De plus la mise en évidence des cycles se heurte à l'effet Slutsky et Yule qui montre que les valeurs obtenues par agrégation suivent des cycles de purs artefacts statistiques.

La méthode des résidus

La méthode des résidus consiste à éliminer des observations brutes ce qui est dû à la tendance séculaire et à la saisonnalité. On admet que ce qui reste – le résidu que l'on ne doit pas confondre avec l'aléatoire – est constitué essentiellement par la fluctuation cyclique. La méthode prend des formes différentes suivant les hypothèses faites quant à la nature et au mode de composition des mouvements.

Dans le cas d'une composition multiplicative, le coefficient cyclique mensuel (respectivement trimestriel) est obtenu à partir de l'équation de base.

$$y_{ik} = T_{ik} + S_k + C_{ik} + Z_{ik}$$

$$C_{ik} = \frac{y_{ik}}{T_{ik} * S_k}$$

Pour éliminer le facteur aléatoire, on procède au lissage des valeurs de C par un calcul des moyennes mobiles.

Pour une composition additive des mouvements, le calcul est analogue.

$$C_{ik} = y_{ik} - T_{ik} - S_{ik}$$

En utilisant les techniques de dessaisonalisation, il est possible d'obtenir des coefficients cycliques C_i relatifs aux années.

La méthode du cycle moyen

Cette méthode a été proposée par le National Bureau of Economic Research des Etats-Unis. Elle part du constat que les fluctuations successives d'une même série ont en général une allure assez voisine pour pouvoir déterminer une oscillation cyclique moyenne.

La méthode utilise des données corrigées des variations saisonnières, mais non du mouvement de tendance générale. La tendance longue est ensuite éliminée au sein de chaque cycle.

Le cycle et subdivisé en plusieurs phases. Une phase d'expansion débute avec le mois de reprise initiale et s'achève au début de la récession. La phase de récession s'engage par le mois de contraction de la phase de cyclique descendante et se termine au début de la reprise.

Chaque cycle est caractérisé par neuf points déterminés de la manière suivante :

1. – mois de reprise initiale ;

2, 3, 4. – la phase d'expansion est subdivisée en trois ;

5. – le premier mois de la récession

6, 7, 8. – la phase de récession est subdivisée par tiers ;

9. – le mois de reprise correspondant au cycle suivant.

Il est possible de calculer la durée moyenne des différentes phases, ainsi que l'amplitude moyenne du cycle pour chacun des neuf points. Cette méthode est intéressante si le cycle se présente comme sensiblement constant en durée et en intensité.

Il sera possible de déterminer un cycle spécifique par produits ou pour les revenus. Ces cycles étant définis, on cherche à construire un cycle global de référence. Dans un second temps, il est possible de comparer chacun des cycles spécifiques au cycle de référence.

L'analyse statistique des chroniques dont nous avons fourni les premiers éléments est d'une grande complexité. Cela est dû aux conditions mêmes de production des données dont nous avons indiqué les difficultés de comparabilité, l'autocorrélation des données (la donnée d'une année dépend des résultats des années précédentes) induit également une autocorrélation des erreurs. Nous n'avons pas supposé également que la composante aléatoire n'intervenait pas dans l'analyse, cette hypothèse, acceptable dans une première approche, devient injustifiée pour des études plus complexes. Un traitement statistique complet des chroniques est évidemment un point essentiel pour la modélisation économétrique en économie appliquée. Les variables économiques évoluent avec le temps, dans leur définition, dans leur extension, le produit intérieur brut augmente avec l'extension des marchés tout autant que par l'expansion de la production. La courbe de vie d'un produit suit une courbe logistique, ce qui conduit parfois à résumer abruptement cet ensemble de phénomènes en affirmant que les séries dépendent du temps.

Chapitre 5

Les indices

Les indices ont été conçus pour effectuer des comparaisons sur des variables économiques mesurables. Ils synthétisent en un seul nombre les modifications affectant un ensemble de variables comme c'est le cas pour l'indice des prix à la consommation. Les changements s'exprimant en valeurs relatives, les comparaisons en sont facilitées. Dans la suite, nous traiterons les variations entre deux dates, les concepts étant facilement transposables aux variations entre deux lieux.

Un indice simple est le rapport des valeurs prises par une grandeur entre deux dates. Un indice synthétique, ou indice composé est un indicateur de tendance centrale d'une distribution d'indices simples. Les indices synthétiques sont souvent des moyennes d'indices simples, qui peuvent être arithmétique ou harmonique.

Le chapitre traitera des indices simples parfois nommés élémentaires pour des produits précisément définis. Les indices synthétiques sont des indicateurs de tendance centrale de distributions d'indices simples fournissant une information sur l'évolution d'une population d'indices. Enfin, la présentation des principes des raccords d'indices permettra d'expliquer comment il est possible d'évaluer des évolutions de grandeurs sur des périodes longues.

Les indices simples

Un indice simple s'applique à une caractéristique clairement définie, le prix d'un produit précis, la quantité d'un article tangible.

Définition

Un indice simple, noté $i(G)_{t/0}$, de la grandeur G est le rapport de la valeur G_t prise par la grandeur à la date 1 à la valeur de G_0 à la date 0, soit :

$$i(G)_{1/0} = \frac{G_1}{G_0}$$

Le calcul d'un indice est une opération de repérage qui suppose d'avoir défini un référentiel : une année de base, une région de référence.

$$i(G)_{1/0} = \frac{G_1}{G_0} \cdot 100$$

Le référentiel de base, choisi en fonction des objectifs des calculs, est affecté du nombre-indice 100 : année base 100, région base 100. Les indices sont dits en base 100 l'année 0 si cette année est celle qui sert de base à la comparaison. Néanmoins, pour les formules de définition et les calculs, nous considérerons des indices de base 1.

Un indice ne mesure pas un niveau en valeur absolue pour une période déterminée, mais en valeur relative. Il n'exprime pas la différence absolue entre G_1 et G_0, il mesure une variation relative de la valeur de G entre deux périodes. Cet indice est un nombre sans dimension. Un nombre-indice cache un pourcentage, en ce sens que la définition d'un indice élémentaire est celle d'un multiplicateur, présenté antérieurement.

Il est possible de calculer des indices élémentaires de prix, de quantité ou de valeur, pour un produit homogène parfaitement défini, c'est le rapport du prix d'une période t sur une période de base 0 pour un produit précisément défini.

$$i(p)_{1/0} = \frac{p_1}{p_0} \text{ ou } i(p)_{1/0} = \frac{p_1}{p_0} \cdot 100 .$$

Au lieu de comparer des prix, il est possible de comparer des quantités, et d'obtenir des indices élémentaires de quantité :

$$i(q)_{1/0} = \frac{q_1}{q_0} \text{ ou } i(q)_{1/0} = \frac{q_1}{q_0} \cdot 100 .$$

Un indice de quantité n'a de sens que pour un produit homogène. Nous pouvons prendre ici l'exemple des cerises bigarreau, celles de taille 22 ne doivent pas être confondues avec les cerises bigarreau de taille 24. Pour les cerises de toutes catégories et tailles, il n'est pas possible de calculer directement un indice de quantité. Cette difficulté concerne également les services pour lesquels le concept de quantité est inapplicable d'où le recours au concept de volume. Les indices de volume résultent généralement de la division d'un indice de valeur par un indice de prix.

$$\text{indice de volume} = \frac{\text{indice de valeur}}{\text{indice de prix}}$$

La notation q est néanmoins utilisée pour les volumes tandis que la notation v est réservée aux valeurs.

Il est aussi possible de calculer un indice élémentaire de valeur par rapport à une année de base. Un indice de valeur s'exprime simplement comme le rapport de deux valeurs :

$$i(v)_{1/0} = \frac{v_1}{v_0} \text{ ou } i(v)_{1/0} = \frac{v_1}{v_0} \cdot 100,$$

d'où nous pouvons donner une écriture plus formalisée de l'indice de volume :

$$i(q)_{1/0} = \frac{i(v)_{1/0}}{i(p)_{1/0}}.$$

Une valeur est le produit du prix du bien considéré par une quantité ou un volume $v = p \times q$.

Il sera alors possible d'écrire un indice élémentaire de valeur comme produit d'un indice de prix et d'un indice de volume.

$$i(v)_{1/0} = \frac{p_1 q_1}{p_0 q_0} = \frac{p_1}{q_1} \cdot \frac{q_1}{q_0} = i(p)_{1/0} \cdot i(q)_{1/0}$$

D'où la propriété fondamentale :

Toute variation d'une valeur se décompose en une variation du volume et une variation du prix.

Indice élémentaire de valeur = indice élémentaire de prix multiplié par l'indice élémentaire de volume. Nous avons là une propriété générale des indices élémentaires.

Par exemple en juin 2013 le prix de la baguette de pain ordinaire était de 0,98 €, il est de 1,05 € en juin 2014. L'indice du prix de la baguette de pain ordinaire est de :

$$i(p)_{juin14/juin13} = \frac{1{,}05}{0{,}98} \cdot 100 \cong 107{,}14$$

Soit une augmentation de 7,14 %, la dépense d'un ménage en baguette de pain ordinaire était en juin 2013 de 29,40 €, elle était de 35,70 € en juin 2014.

L'indice de valeur, ici l'indice de la dépense en baguette de pain ordinaire, est donc

$$i(v)_{juin14/juin13} = \frac{35{,}7}{29{,}4} \cdot 100 \cong 121{,}43$$

Il est possible de calculer l'indice de volume

$$i(q)_{juin14/juin13} = \frac{121{,}43}{107{,}14} \cdot 100 \cong 113{,}3$$

La variation de la dépense en baguette de pain ordinaire s'explique par une augmentation des prix de 7,14 % et du volume de 13,3 %.

Propriétés des indices élémentaires

Les indices élémentaires possèdent trois propriétés : l'identité, la circularité, la réversibilité.

L'identité

L'identité signifie que $i_{t/t} = 100$ ou $i_{0/0} = 100$. ; $i(G)_{0/0} = \frac{G_0}{G_0} = 1$, elle est donc vérifiée.

$$i(p)_{juin14/juin14} = \frac{0,98}{0,98} \cdot 100 = 100$$

L'identité permet de choisir une année de référence c'est-à-dire définir une base 100. Cette année de base constituera la référence à partir de laquelle se feront les comparaisons.

La réversibilité

La propriété de réversibilité s'exprime de la manière suivante : $i(G)_{0/1} = \frac{1}{i(G)_{1/0}}$, elle est vérifiée comme suit

$$i(G)_{0/1} = \frac{G_0}{G_1} = \frac{1}{G_1 / G_0} = \frac{1}{i(G)_{1/0}}$$

$$i(p)_{juin13/juin14} = \frac{1}{1,074} \cdot 100 = 93,3$$

En base 100 juin 2014, l'indice du prix du pain était de 93,3 en juin 2013 (6,7 % moins cher). La période de base peut être modifiée grâce à la propriété de réversibilité.

La circularité

La propriété de circularité signifie que, quel que soit le cheminement suivi, l'indice garde la même valeur. Elle se traduit par l'expression :

$i(G)_{t/0} = i(G)_{t/t'} \cdot i(G)_{t'/0}$.

Cette propriété est vérifiée, en effet, $i(G)_{t/0} = \frac{G_t}{G_0} = \frac{G_t}{G_{t'}} = \frac{G_{t'}}{G_0} = i(G)_{t/t'} \cdot i(G)_{t'/0}$.

Si nous supposons que le prix de la baguette de pain ordinaire en juin 2012 était de 0,95 €, nous pouvons calculer l'indice du prix du pain en base 100 juin 2012 :

$$i(p)_{juin13/juin12=} = \frac{0,98}{0,95} \cdot 100 \cong 103,16$$

$$i(p)_{juin14/juin12} = i(p)_{juin14/juin13} \cdot i(p)_{juin13/juin12}$$

$$i(p)_{juin14/juin12} = 1,0714 \cdot 1,0316 \cdot 100 \cong 110,53 .$$

Ce résultat signifie que le prix a augmenté de 10,53 %

Cette propriété permet de calculer l'indice de l'année t par rapport à l'année t' si l'on connaît les indices des années t et t' par rapport à une même base 0.

$$i(G)_{t/t'} = \frac{i(G)_{t/0}}{i(G)_{t'/0}} \text{ ou } i(G)_{t/t'} = \frac{i(G)_{t/0}}{i(G)_{t'/0}} \cdot 100$$

Il sera possible de modifier l'année de référence sans avoir à effectuer tous les calculs.
Dans le cas de trois périodes, trois écritures sont possibles :

$$i(G)_{3/0} = i(G)_{3/2} \cdot i(G)_{2/1} \cdot i(G)_{1/0} = i(G)_{3/1} \cdot i(G)_{1/0} = i(G)_{3/2} \cdot i(G)_{2/0}$$

Cette propriété se généralise pour un nombre quelconque de périodes :

$$i(G)_{t/0} = i(G)_{t/t-1} \cdot i(G)_{t-1/t-2} \cdots i(G)_{1/0} = \prod_{j=1}^{t} i(G)_{j/j-1}$$

Cette formule définit une chaîne d'indices.

Nous pouvons utiliser l'exemple de l'évolution du prix d'un Smartphone.

Tableau 1. Évolution des prix d'un Smartphone (données fictives).

Années	Prix €	Ventes €	Indices prix	Indices de valeurs	Indices de volume
	p_i	v_i	$i(p)_{t/t-1} = \frac{p_t}{p_{t-1}} \cdot 10$	$i(v)_{t/t-1} = \frac{v_t}{v_{t-1}} \cdot 100$	$i(q)_{t/t-1} = \frac{i(v)_{t/t-1}}{i(p)t /_{t-1}} \cdot 100$
2012	345	34 500			
2013	250	35 000	72,46	101,45	140,0
2014	220	33 000	88,00	94,28	107,14
Indices 2014/ 2012			63,8	95,65	150,0

Une diminution des prix de 36,2 % (100-63,8) se conjugue avec une diminution des ventes de 4,35 ce qui se traduit par une augmentation de 50 % du volume.

L'élasticité prix se calcule ainsi : .

$$e_p = \frac{\frac{\Delta q}{q}}{\frac{\Delta p}{p}} = \frac{\Delta q}{\Delta p} \cdot \frac{p}{q} = \frac{\Delta p}{q} \cdot \frac{p}{\Delta p} = \frac{150}{63{,}8} \cong 2{,}35$$

L'augmentation des prix n'a pas réduit le volume des ventes du produit bien au contraire.

Les indices élémentaires de prix avec changement de base

Disposant du tableau suivant des indices élémentaires de prix en base 100 l'année précédente, l'objectif est d'obtenir les indices de prix en base 100 l'année 0.

Tableau 2. Indices base 100 année 0.

	Année 0	Année 1	Année 2	Année 3
A	101	103	101	104
B	102	104	100	105
C	103	102	102	103
D	102	102	104	105

La propriété de circularité des indices simples permet de répondre à la question. Dans le cas des indices simples l'indice de l'année 3 en base 100 l'année 0 peut s'écrire :

$i(p)^h_{3/0} = i(p)^h_{3/2} \cdot i(p)^h_{2/1} \cdot i(p)^h_{1/0}$.

Pour faciliter la compréhension des calculs, le tableau ci-dessous explicite la signification des données.

Tableau 3. Indices des prix en base 100 année précédente.

	Année 0	Année 1	Année 2	Année 3
Indices	$i(p)^h_{0/-1}$	$i(p)^h_{1/0}$	$i(p)^h_{2/1}$	$i(p)^h_{3/2}$
A	101	103	101	104
B	102	104	100	105
C	103	102	102	103
D	102	102	104	105

Pour le produit A, l'application de cette propriété permet d'obtenir les indices suivants : $i(p)^h_{0/0} = 100$ (propriété d'identité), l'indice $i(p)^A_{1/0}$ est connu, il est fourni, en l'occurrence 103. L'indice $i(p)^A_{2/0}$ est obtenu en appliquant la propriété de circularité :

$i(p)^A_{2/0} = i(p)^A_{1/0} \cdot i(p)^A_{2/1} = 1.03 \cdot 1.01 \cdot 100 = 104.0$.

De même pour l'indice $i(p)^A_{3/0}$:

$i(p)^A_{3/0} = i(p)^A_{1/0} \cdot i(p)^A_{2/1} \cdot i(p)^A_{3/2} = 1.03 \cdot 1.01 \cdot 1.04 \cdot 100 = 108.2$.

Le tableau des indices en base 100 année 0 est le suivant :

Tableau 4. Indices des prix en base 100 année 0.

$i(p)^h_{t/0}$	Année 0	Année 1	Année 2	Année 3
A	100,0	103,0	104,0	108,2
B	100,0	104,0	104,0	109,2
C	100,0	102,0	104,0	107,2
D	100,0	102,0	106,1	111,4

Si les indices de prix pour chaque produit sont calculés grâce aux propriétés des indices simples, l'évolution de l'indice global des prix pour chaque année ne peut être calculée. Ce sera l'objet des paragraphes suivants.

Exemple : indices élémentaires de valeur base 100 année 0

À partir des indices des valeurs des consommations pour les années 1, 2 et de différents produits et de la consommation totale en base 100 année 0, il est possible de calculer les indices élémentaires de valeur.

Tableau 5. Structure des dépenses par fonction de consommation (milliers d'euros).

Consommations	Année 0	Année 1	Année 2	Année 3
A	400	470	480	470
B	360	370	380	350
C	120	160	170	160
D	240	240	250	230
Total	1 120	1 240	1 280	1 210

Il est alors possible de calculer les indices de valeur pour chaque catégorie de produits et pour l'ensemble avec des données arrondies :

$$i(v)^h_{t/0} = \frac{v^h_t}{v^h_0} \cdot 100 \cdot$$

Par exemple pour le produit B pour l'année 2, $i(v)^B_{2/0} = \frac{v^B_2}{v^B_0} = \frac{380}{360} \cdot 100 = 105{,}6$.

Les indices de valeur pour chaque année sont calculés directement en référence à la somme des valeurs de l'année 0 :

$$i(v)_{3/0} = \frac{v_3}{v_0} = \frac{1210}{1120} \cdot 100 = 108{,}0 \cdot$$

Tableau 6. Indices valeurs base 100 année 0.

	Année 0	Année 1	Année 2	Année 3
A	100,0	117,5	120,0	117,5
B	100,0	102,8	105,6	97,2
C	100,0	133,3	141,7	133,3
D	100,0	100,0	104,2	95,8
Indices de valeur globale	100,0	110,7	114,3	108,0

Il est alors possible d'obtenir le tableau des indices de volume pour chacun des produits et chacune des années, en utilisant la relation entre les indices de valeur et les indices de prix :

$$i(q)^h_{t/0} = \frac{i(v)^h_{t/0}}{i(p)^h_{t/0}},$$

$$i(q)^A_{2/0} = \frac{i(v)^A_{2/0}}{i(p)^A_{2/0}} = \frac{105,6}{104,0} \cong 101,5 \cdot$$

Tableau 7. Indices des volumes (résultats arrondis).

	Année 0	Année 1	Année 2	Année 3
A	100,0	114,1	115,4	108,6
B	100,0	98,8	101,5	89,0
C	100,0	130,7	136,2	124,4
D	100,0	98,0	98,2	86,0

Il est possible de suivre l'évolution des volumes pour chaque produit. Il n'est cependant pas possible de disposer d'un indice de volume global pour chaque année faute d'une formule permettant d'obtenir une tendance centrale des indices. Les indices synthétiques répondent à cette difficulté.

Exemple : indices des voyageurs

Les indices peuvent être utilisés directement pour suivre l'évolution d'une grandeur, ici un indice de quantité pour des voyageurs. Soit la distribution suivante du nombre de voyageurs (en millions) transportés en métro par trimestres au cours de quatre années consécutives :

Tableau 8. Données de base.

Trimestres / Années	I	II	III	IV
1	48	44	40	44
2	56	52	44	48
3	64	52	44	56
4	72	68	48	72

Source : RATP

Il est possible de calculer en base 100 le premier trimestre de l'année 1 l'indice du nombre de passagers. Ici, il s'agit bien de quantité et non de volume, car nous disposons du nombre de voyageurs, quel que soit le prix

qu'ils ont acquitté pour utiliser le métro. Nous avons à calculer les indices du nombre de passagers en prenant pour période de référence le premier trimestre de l'année 1.
La définition générale d'un indice est : $i_{ik/1I} = \frac{G_{ik}}{G_{1I}}$.

Ici G_{1I} est le nombre de passagers pour le premier trimestre de l'année 1 soit 48. G_{ik} correspond au nombre de passagers transportés l'année i pour le trimestre k.
L'indice pour le second trimestre de l'année 1 est donc : $i_{1II/1I} = \frac{44}{48} \cdot 100 \cong 91{,}7$
ou encore, $i_{3II/1I} = \frac{G_{3II}}{G_{1I}} = \frac{52}{48} \cdot 100 \cong 108{,}3$.

Nous obtenons alors le tableau des indices simples, il s'agit d'indices de quantité.

Tableau 9. Indices base 100 =1,I.

Trimestres / Années	I	II	III	IV
1	100,0	91,7	83,3	91,7
2	116,7	108,3	91,7	100,0
3	133,3	108,3	91,7	116,7
4	150,0	141,7	100,0	150,0

Il est envisageable de faire les calculs pour un indice base 100 année 2. La base de référence est le nombre moyen de passagers par trimestre au cours de l'année 2. La moyenne arithmétique du nombre de passagers de l'année est de 50. Nous allons maintenant calculer les indices de chaque trimestre en prenant pour base cette moyenne, selon la formule suivante
$i_{ik} = \frac{G_{ik}}{50} \cdot 100$.

L'indice du troisième trimestre de la deuxième année sera donc :
$i_{2III} = \frac{44}{50} \cdot 100 = 88$.

Tableau 10. Indices 100 année 2.

Trimestres / Années	I	II	III	IV
1	96,0	88,0	80,0	88,0
2	112,0	104,0	88,0	96,0
3	128,0	104,0	88,0	112,0
4	144,0	136,0	96,0	144,0

Pour déterminer les indices en base 100 pour l'ensemble des trimestres, la démarche est identique à celle de la question précédente. Il suffit de calculer le nombre moyen de passagers par trimestre, soit 53,25, puis de calculer les différents indices. Ce chiffre est la moyenne arithmétique du nombre trimestriel de passagers au cours des quatre années.

La forme générale de l'indice sera donc : $i_{ik} = \frac{G_{ik}}{53,25} \cdot 100$.

Nous calculons comme exemple l'indice relatif au quatrième trimestre de la quatrième année :

$$i_{4IV} = \frac{72}{53,25} \cdot 100 \cong 135,2 \ .$$

Tableau 11. Indices 100 moyenne générale.

Trimestres / Années	I	II	III	IV
1	90,1	82,6	75,1	82,6
2	105,2	97,7	82,6	90,1
3	120,2	97,7	82,6	10,5,2
4	135,2	127,7	90,1	135,2

Ce tableau précise le caractère saisonnier du nombre de voyageurs transportés, il peut servir à des prévisions pour l'organisation des transports.

Les indices synthétiques

Il est rare de calculer l'évolution du prix d'un seul bien, habituellement, on calcule plutôt un indicateur global de l'évolution des prix pour un ensemble de produits. Un indice synthétique, ou indice composé, est un indicateur de tendance centrale d'une distribution d'indices simples, appelés parfois indices élémentaires. Les indices synthétiques sont des moyennes d'indices simples : moyenne arithmétique pour l'indice de Laspeyres (Étienne Laspeyres économiste allemand 1834-1913), moyenne harmonique pour l'indice de Paasche (Hermann Paasche. statisticien allemand 1851-1925).
Nous insisterons sur les indices de prix, d'une part les définitions étant facilement transposables aux indices de volume, d'autre part il est exceptionnel dans le domaine de l'économie de pouvoir calculer des indices de volume directement sans passer par les prix (problème de la sommation de quantités exprimées dans des unités différentes).

Les indices de Laspeyres

La moyenne arithmétique apparaît naturellement comme la solution la plus simple pour synthétiser une distribution d'indices élémentaires. L'indice synthétique de Laspeyres des prix est une moyenne arithmétique des indices élémentaires des prix.

La pondération pour un produit donné est son importance relative dans le panier de consommation ; ce que nous avons défini, auparavant, comme la fréquence. Les pondérations retenues sont celles de la période de base. La structure de la période de base sert de référence pour évaluer les évolutions, les transformations de la structure n'interviennent pas dans les calculs. Cette hypothèse suppose que les agents économiques pourraient ne pas modifier les volumes demandés face à une variation des prix, en opposition aux résultats de l'analyse économique. Les indices de Laspeyres comparent l'année courante, la situation présente, à l'année de référence passée.

La formule de l'indice de Laspeyres des prix a la forme générale d'une moyenne arithmétique.

Soit $i(p)_{1/0}^h = \frac{p_1^h}{p_0^h}$ l'indice élémentaire du prix du produit h ;

k_0^h représente le coefficient budgétaire de l'année ou la période de base, l'importance relative de la dépense en produit h dans le total des dépenses de consommation à l'année 0.

Soit $V_0^h = p_0^h q_0^h$ la valeur élémentaire de la dépense l'année zéro en produit h. La somme de ces dépenses élémentaires constitue la dépense totale pour l'année 0 :

$$V_0 = \sum_{h=1}^{n} p_0^h q_0^h .$$

L'importance relative de la dépense en produit h, sa fréquence relative, est donc :

$$k_0^h = \frac{V_0^h}{V_0} = \frac{p_0^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} .$$

Avec $i(p)_{1/0}^h$ l'indice élémentaire du prix du bien h, la formule de l'indice de Laspeyres est alors :

$$L(p)_{1/0} = \sum_{h=1}^{n} k_0^h \cdot i(p)_{1/0}^h .$$

La formule ci-dessus est celle à utiliser pour les calculs. L'indice de Laspeyres des prix suppose que les volumes sont ceux de l'année de base et que seuls les prix changent comme l'indique la formule théorique suivante :

$$L(p)_{1/0} = \sum_{h=1}^{n} k_h^0 i(p)_{1/0}^h = \sum_{h=1}^{n} \frac{p_h^0 q_h^0}{\sum_{h=1}^{n} p_h^0 q_h^0} \times \frac{p_h^1}{p_h^0} .$$

Cette formule se transforme en :

$$L(p)_{1/0} = \sum_{h=1}^{n} \frac{p_0^h q_0^h \dfrac{p_1^h}{p_0^h}}{\sum_{h=1}^{n} p_0^h q_0^h} = \frac{\sum_{h=1}^{n} p_1^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} \cdot$$

Cette expression montre que l'indice de Laspeyres des prix compare la situation de l'année 1 à celle de l'année de base.
L'indice de Laspeyres des volumes se définit de façon analogue en utilisant les indices élémentaires de volume. Laspeyres des volumes $L(q)$ s'exprime comme la moyenne arithmétique des indices élémentaires des volumes pondérés par les mêmes coefficients.

$$L(q)_{1/0} = \sum_{h=1}^{n} k_0^h i(q)_{1/0}^h$$

$$L(q)_{1/0} = \sum_{i=1}^{n} \frac{p_0^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} \cdot \frac{q_1^h}{q_0^h} = \frac{\sum_{i=1}^{n} p_0^h q_1^h}{\sum_{h=1}^{n} p_0^h q_0^h}$$

Les indices de Laspeyres possèdent la propriété d'identité, mais ni la propriété de réversibilité, ni celle de circularité.

On vérifie aisément que l'indice de Laspeyres vérifie la propriété d'identité qui s'écrit :

$$L(p)_{0/0} = 1$$

$$L(p)_{0/0} = \frac{\sum_{h=1}^{n} p_h^0 q_h^0}{\sum_{h=1}^{n} p_h^0 q_h^0} = 1$$

Si l'indice de Laspeyres était réversible alors la relation $L(p)_{0/1} = \dfrac{1}{L(p)_{1/0}}$ serait vérifiée.

$$\frac{1}{L(p)_{1/0}} = \frac{\sum_{h=1}^{n} p_0^h q_0^h}{\sum_{h=1}^{n} p_1^h q_0^h} \text{ et } L(p)_{0/1} = \frac{\sum_{h=1}^{n} p_h^0 q_h^1}{\sum_{h=1}^{n} p_h^1 q_h^1}$$

Nous pouvons donc en conclure que l'indice est clairement non réversible.
La propriété de circularité s'exprime par la relation suivante :

$$L(p)_{2/0} = L(p)_{2/1} \cdot L(p)_{1/0}$$

$$L(p)_{1/0} = \sum_{h=1}^{n} k_h^0 i(p)_{1/0}^h$$

$$L(p)_{1/0}=\frac{\sum_{h=1}^{n}p_1^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}$$

$$L(p)_{2/1}=\frac{\sum_{h=1}^{n}p_2^h q_1^h}{\sum_{h=1}^{n}p_1^h q_1^h}$$

$$L(p)_{2/1}\cdot L(p)_{1/0}=\frac{\sum_{h=1}^{n}p_1^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}\cdot\frac{\sum_{h=1}^{n}p_2^h q_1^h}{\sum_{h=1}^{n}p_1^h q_1^h}.$$

Le résultat obtenu n'est manifestement pas l'indice de Laspeyres attendu :

$$L(p)_{2/0}=\frac{\sum_{h=1}^{n}p_2^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}.$$

La division d'un indice de Laspeyres par un autre indice de Laspeyres conduit au phénomène des pondérations implicites.

Les problèmes de comparaisons : les pondérations implicites

L'usage fréquent des indices de prix et de volume conduit à comparer des évolutions par rapport à une période qui n'est pas la période de base.
L'indice utilisé est un indice de Laspeyres donc :

$$L(p)_{2/0}=\frac{\sum_{h=1}^{n}p_2^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}\ \text{ et }\ L(p)_{1/0}=\frac{\sum_{h=1}^{n}p_1^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}.$$

Le rapport ci-dessus se calcule comme le rapport de $L(p)_{2/0}$ sur $L(p)_{1/0}$:

$$\frac{L(p)_{2/0}}{L(p)_{1/0}}=\frac{\dfrac{\sum_{h=1}^{n}p_2^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}}{\dfrac{\sum_{h=1}^{n}p_1^h q_0^h}{\sum_{h=1}^{n}p_0^h q_0^h}}=\frac{\sum_{h=1}^{n}p_2^h q_0^h}{\sum_{h=1}^{n}p_1^h q_0^h}=\sum_{h=1}^{n}\frac{p_1^h q_0^h}{\sum_{h=1}^{n}p_1^h q_0^h}\times\frac{p_2^h}{p_1^h}.$$

L'indice de Laspeyres des prix de la période 2 par rapport à la période 1 serait le suivant :

$$L(p)_{2/1} = \frac{\sum_{h=1}^{n} p_2^h q_1^h}{\sum_{h=1}^{n} p_1^h q_1^h} = \sum_{h=1}^{n} \frac{p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_1^h} \times \frac{p_2^h}{p_1^h} \quad .$$

Le rapport des indices pondère les indices élémentaires de prix par $\frac{p_1^h q_0^h}{\sum_{h=1}^{n} p_1^h q_0^h}$. Les coefficients ainsi obtenus sont appelés coefficients de pondération implicites.

Cette pondération diffère de la pondération théorique qui est $\frac{p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_1^h}$ comme de la pondération de base qui est $\frac{p_0^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h}$.

Ils se déduisent des coefficients de pondération d'origine. Ils sont multipliés par $\frac{p_1^h}{p_0^h}$ et divisés par

$$L(p)_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} \, .$$

$$\frac{p_1^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} = \frac{p_0^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} * \frac{p_1^h}{p_0^h} * \frac{\sum_{h=1}^{n} p_0^h q_0^h}{\sum_{h=1}^{n} p_1^h q_0^h} = k_0^h * i(p)_{1/0}^h * \frac{1}{L(p)_{1/0}}$$

La comparaison de deux indices augmente la pondération des produits dont les prix ont augmenté plus que l'indice général et réduit la pondération dans le cas inverse.
Le produit de deux indices ou la division d'un indice par un autre ne donne pas un indice de Laspeyres. Le calcul des indices intermédiaires nécessite de recalculer les pondérations pour chaque année de base.
Par contre, en application des propriétés de la moyenne, l'indice de Laspeyres dispose de la propriété d'agrégation. Il sera possible de calculer les indices pour des sous-populations puis d'obtenir l'indice des prix pour toute la population étudiée.

Exemple : calcul d'un indice de Laspeyres

Les données sont identiques à celles utilisées dans l'exemple numérique des indices simples.

Tableau 12. Indices base 100 année 0.

$i(p)^h_{i/0}$	Année 0	Année 1	Année 2	Année 3
A	100,0	103,0	104,0	108,2
B	100,0	104,0	104,0	109,2
C	100,0	102,0	104,0	107,2
D	100,0	102,0	106,1	111,4

Tableau 13. Structure des dépenses par fonction de consommation (milliers d'euros).

Valeurs	Année 0	Année 1	Année 2	Année 3
A	400	470	480	470
B	360	370	380	350
C	120	160	170	160
D	240	240	250	230
Total	1 120	1 240	1 280	1 210

Tableau 14. Structure des dépenses par fonction de consommation (en %).

	Année 0	Année 1	Année 2	Année 3
A	35,7	37,9	37,5	38,8
B	32,1	29,8	29,7	28,9
C	10,7	12,9	13,3	13,2
D	21,4	19,4	19,5	19,0
Total	100,0	100,0	100,0	100,0

Tableau 15. Indice de Laspeyres des prix.

Produits	k_0^h	$i(p)^h_{3/0}$	$k_0^h \cdot i(p)^h_{3/0}$
A	35,7	108,2	3 863,971429
B	32,1	109,2	3 510
C	10,7	107,2	1 148,155714
D	21,4	111,4	2 386,8
	100,0		10 908,92714

$$L(p)_{3/0} = \sum_{h=1}^{n} k_0^h \cdot i(p)^h_{3/0} = \frac{10908,92714}{100} \cong 109,09$$

Le calcul de l'indice permet d'identifier une augmentation de 9,09 % des prix. Il peut sembler logique de chercher un indice de volume en divisant l'indice de valeur par l'indice de Laspeyres des prix. Le résultat sera un indice de volume, qui ne peut être un indice de Laspeyres puisque le produit de deux indices de Laspeyres n'est pas un indice de Laspeyres. Le paragraphe suivant permettra de le définir.

L'analyse d'un indice de valeur

Une valeur économique V est le produit de prix et de quantité ; dans le cas des indices simples, la relation était particulièrement aisée entre les indices de prix et de quantité. Dans le cas des indices synthétiques, l'indice de valeur est-il le produit d'un indice de Laspeyres des prix par un indice de Laspeyres des volumes ? Ce n'est pas aussi évident, même si un indice de valeur est toujours le produit d'un indice de prix par un indice de volume.

Soit $V = \sum_{h=1}^{n} p_h q_h$ la somme des valeurs élémentaires. Si V_0 est la valeur de la période de base et V_1 celle de la période terminale, alors :

$$V_0 = \sum_{h=1}^{n} p_0^h q_0^h \text{ et } V_1 = \sum_{h=1}^{n} p_1^h q_1^h .$$

L'indice $v_{1/0}$, indice de valeur, est le rapport des deux valeurs $i(v)_{1/0} = \frac{V_1}{V_0} = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{V \sum_{h=1}^{n} p_0^h q_0^h}$.

Un indice de cette sorte n'a pas d'interprétation immédiate puisqu'il dépend des modifications des prix et des volumes sans que l'on sache ce qui provient des prix et ce qui revient aux quantités. L'indice de Laspeyres est

$$L(p)_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h},$$

$$\frac{v_{1/0}}{L(p)_{1/0}} = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{V \sum_{h=1}^{n} p_0^h q_0^h} \cdot \frac{\sum_{h=1}^{n} p_0^h q_0^h}{\sum_{h=1}^{n} p_1^h q_0^h} = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_0^h} .$$

Pour obtenir l'indice de valeur $v_{1/0}$ il faut multiplier l'indice de prix par le rapport :

$$I(q) = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_0^h} .$$

Il s'agit d'un indice de volume, puisque les prix sont ceux de l'année par contre les volumes sont ceux des années 1 et 0, cet indice n'est donc pas un indice de Laspeyres. Il compare la valeur de l'année l à ce qu'aurait été cette valeur si les volumes étaient ceux de l'année 0 et les prix ceux de l'année 1. L'indice de valeur s'écrit comme produit d'un indice de prix et d'un indice de volume.

$$i(v)_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} \cdot \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_0^h}$$

L'indice de valeur est donc bien le produit d'un indice de prix multiplié par un indice de volume.
Il est également possible de chercher quel indice de prix multiplié par l'indice de Laspeyres des volumes donne l'indice de valeur.

$$L(q)_{1/0} = \frac{\sum_{i=1}^{n} p_0^h q_1^h}{\sum_{h=1}^{n} p_0^h q_0^h}$$

Pour obtenir l'indice de valeur, il faut multiplier l'indice de Laspeyres des volumes par le rapport

$$\frac{\sum_{h=1}^{n} p_h^1 q_h^1}{\sum_{h=1}^{n} p_h^0 q_h^1},$$

qui est un indice de prix. D'où une seconde décomposition de l'indice de valeur :

$$v_{1/0} = \frac{\sum_{h=1}^{n} p_0^h q_1^h}{\sum_{h=1}^{n} p_0^h q_0^h} \cdot \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_0^h q_1^h}.$$

L'indice de valeur peut s'écrire de deux manières différentes comme le produit d'un indice de prix et d'un indice de volume.

$$v_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_0^h}{\sum_{h=1}^{n} p_0^h q_0^h} \cdot \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_0^h} = \frac{\sum_{h=1}^{n} p_0^h q_1^h}{\sum_{h=1}^{n} p_0^h q_0^h} \cdot \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_0^h q_1^h}$$

L'indice des prix et un indice de Paasche, il s'écrit comme suit :

$$P(p)_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_0^h q_1^h}$$

tandis que l'indice de Paasche des volumes est formulé par :

$$P(q)_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_0^h}.$$

L'équation fondamentale liant les indices de prix, de volume et valeur est respectée pour les indices synthétiques.

$$i(v)_{1/0} = L(p)_{1/0} \cdot P(q)_{1/0} = L(q)_{1/0} \cdot P(p)_{1/0}$$

L'objet du paragraphe suivant permettra de donner aux indices de Paasche une forme simple.

Les indices de Paasche

L'indice de Paasche des prix est une moyenne harmonique des indices élémentaires des prix. La pondération retenue est, dans ce cas, celle de l'année terminale. L'indice de Paasche des prix est la moyenne harmonique des indices élémentaires des prix pondérés par les coefficients de l'année courante :

$$\frac{1}{P(p)_{1/0}} = \sum_{h=1}^{n} \frac{k_1^h}{i(p)_{1/0}^h} \quad \text{où} \quad k_1^h = \frac{V_1^h}{V_1} = \frac{p_1^h q_1^h}{\sum_{h=1}^{n} p_1^h q_1^h}$$

représente le coefficient budgétaire à l'époque de base et $i(p)_{1/0}^h$ l'indice élémentaire du prix du bien i.

L'indice de Paasche des volumes se définit simplement comme la moyenne harmonique des indices élémentaires de volume.

$$\frac{1}{P(q)_{1/0}} = \sum_{h=1}^{n} \frac{k_1^h}{i(q)_{1/0}^h}$$

Il est possible de donner une forme plus théorique à l'indice de Paasche :

$$P(p)_{1/0} = \frac{\sum_{h=1}^{n} p_1^h q_1^h}{\sum_{h=1}^{n} p_0^h q_1^h}$$

L'indice de Paasche revient à comparer la période 1 à la période 0 en supposant que la situation de référence est celle de l'année 1. Il compare donc le passé au présent. Il possède la propriété d'identité et d'agrégation, mais ni celle de réversibilité ni celle de circularité. L'utilisation de cet indice se heurte aux mêmes difficultés que pour l'indice de Laspeyres et les explications fournies pour l'indice de Laspeyres sont analogues pour les indices de Paasche.

Exemple

Tableau 16. Indices base 100 année 0.

$i(p)^h_{i/0}$	**Année 0**	**Année 1**	**Année 2**	**Année 3**
A	100,0	103,0	104,0	108,2
B	100,0	104,0	104,0	109,2
C	100,0	102,0	104,0	107,2
D	100,0	102,0	106,1	111,4

Tableau 17. Les coefficients budgétaires.

	k^h_0	k^h_1	k^h_2	k^h_3
A	35,7	37,9	37,5	38,8
B	32,1	29,8	29,7	28,9
C	10,7	12,9	13,3	13,2
D	21,4	19,4	19,5	19,0
Total	100,0	1 247	100,0	100,0

Tableau 18. Indice de Paasche des prix.

Produits	k^h_3	$i(p)^h_{3/0}$	$\frac{k^h_3}{i(p)^h_{3/0}}$
A	38,8	108,2	0,359021577
B	28,9	109,2	0,264886629
C	13,2	107,2	0,123394853
D	19,0	111,4	0,170655251
	100,0		0,917958309

$$\frac{1}{P(p)_{3/0}} = \sum_{h=1}^{n} \frac{k^h_3}{i(p)^h_{3/0}} = \frac{0{,}917958309}{100} = 0{,}00917958309$$

$$P(p)_{3/0} = \frac{1}{0{,}00917958309} \cong 108{,}94$$

Soit une augmentation de 8,94 % des prix.

L'indice de volume correspondant à un indice de Laspeyres des volumes :

$$L(q)_{3/0} = \frac{v_{3/0}}{P(p)_{3/0}} = \frac{108}{108,94} \cong 99,17$$

Nous obtenons deux résultats différents pour l'évaluation de la hausse des prix, même si l'écart est minime. Nous n'avons aucun moyen de choisir entre ces deux informations tout aussi pertinentes l'une que l'autre. Le choix de l'indice est, en dehors des contraintes sur les données, dépend de la vision retenue.

L'indice de Fisher

Devant cette difficulté, un mathématicien américain, Irving Fisher (1867-1947) a proposé une autre série d'indices. C'est un indice synthétique de second niveau puisqu'il est la moyenne d'indices synthétiques. L'indice de Fisher est la moyenne géométrique des indices de Laspeyres et de Paasche.

Pour les prix il se calcule avec $F(p)_{1/0} = \sqrt{L(p)_{1/0} \cdot P(p)_{1/0}} = \left[L(p)_{1/0} \cdot P(p)_{1/0}\right]^{\frac{1}{2}}$ et pour les volumes

$$F(q)_{1/0} = \sqrt{L(q)_{1/0} \cdot P(q)_{1/0}} = \left[L(q)_{1/0} \cdot P(q)_{1/0}\right]^{\frac{1}{2}}$$

L'indice de Fisher possède, outre la propriété d'identité, la propriété de réversibilité :

$$F(p)_{0/1} = \frac{1}{F(p)_{1/0}}$$

Il ne possède par contre pas la propriété de circularité et son interprétation n'est pas aisée.
Dans notre exemple, l'indice de Fisher se calcule facilement :

$$F(p)_{3/0} = \sqrt{L(p)_{3/0} \cdot P(p)_{3/0}} = \sqrt{109,09 \cdot 108,94} = 109,01$$

La hausse est cette fois de 9,01 % ; nous disposons maintenant de trois réponses pour l'évolution des prix. Par application de l'équation fondamentale, il est possible d'obtenir les indices de volume.

$$v_{1/0} = L(p)_{1/0} \cdot P(q)_{1/0} = L(q)_{1/0} \cdot P(p)_{1/0}$$

Le rapport $\frac{V_{1/0}}{L(p)_{1/0}}$ permet d'obtenir l'indice de volume correspondant $P(q)_{1/0}$, qui est un indice Paasche :

$$P(q)_{3/0} = \frac{v_{3/0}}{L(p)_{3/0}} = \frac{108,0}{109,09} \cong 99,03$$

De même, le rapport $\frac{V_{1/0}}{P(p)_{1/0}}$ permet d'obtenir l'indice de volume, qui est un indice de Laspeyres :

$$L(q)_{3/0} = \frac{v_{3/0}}{P(p)_{3/0}} = \frac{108,0}{108,94} \cong 99,17$$

Il est également possible de calculer l'indice de Fisher des volumes :

$$F(q)_{3/0} = \frac{v_{3/0}}{F(p)_{3/0}} = \frac{108,0}{109,01} \cong 99,1$$

L'évaluation de l'évolution des prix et des volumes dépend des formules retenues. Une quatrième manière de calculer une évolution des prix utilise la formule des indices chaînes. Pour des raisons pratiques, la grande majorité des indices de prix utilise une formule de Laspeyres. L'inconvénient majeur de ce choix est de ne pas intégrer les changements des structures.

Les indices chaînes

Les indices vieillissent, l'apparition de nouveaux produits, les évolutions des comportements ou des consommations modifient la structure du phénomène à étudier. Les pondérations initiales ne correspondent plus exactement à la réalité du moment, ce qui conduit à des biais systématiques difficiles à apprécier. L'utilisation fréquente des indices, en particulier ceux de prix, impose d'estimer les évolutions par rapport à une période qui n'est pas toujours la période de base. Ces comparaisons utilisent, sans le savoir, des pondérations implicites dont la signification est peu claire.

Les indices chaînes répondent à ces besoins de réajustements permanents. Un indice chaîne est le produit d'indices successifs. Le nouvel indice est obtenu à partir des indices synthétiques calculés pour chacune des années. L'indice chaîne se définit par :

$$C(p)_{t/0} = I(p)_{t/t-1} \cdot I(p)_{t-1/t-2} \cdot \ldots \cdot I(p)_{2/1} \cdot I(p)_{1/0}$$

Les chaînons sont le plus souvent des indices de type Laspeyres.
La formule générale d'un indice chaîne dont les chaînons sont des indices de Laspeyres s'écrit comme suit.

$$C(p)_{t/0} = L(p)_{t/t-1} \cdot L(p)_{t-1/t-2} \cdot \ldots \cdot L(p)_{2/1} \cdot L(p)_{1/0}$$

Chaque année, l'indice de Laspeyres des prix est calculé sur une base 100 l'année précédente. Cette méthode permet de modifier chaque année les pondérations et donc de prendre compte des modifications de la structure de la consommation. On calcule donc la série d'indices :

$$L(p)_{1/0} = \sum_{h=1}^{n} k_0^h \cdot i(p)_{1/0}^h \text{ ; } L(p)_{2/1} = \sum_{h=1}^{n} k_1^h \cdot i(p)_{2/1}^h \text{ ;........ ; } L(p)_{t/t-1} = \sum_{h=1}^{n} k_{t-1}^h \cdot i(p)_{t/t-1}^h .$$

L'indice chaîne $C(p)_{2/0}$ est donc égal à :

$$C(p)_{2/0} = L(p)_{1/0} * L(p)_{2/1} = \sum_{h=1}^{n} k_h^0 i(p)_{1/0}^h * \sum_{h=1}^{n} k_h^1 i(p)_{2/1}^h .$$

L'indice $C(p)_{t/0}$ s'obtient très facilement :

$$C(p)_{t/0} = L(p)_{t/t-1} \cdot ...... \cdot L(p)_{2/1} \cdot L(p)_{1/0} .$$

L'indice chaîne possède à l'évidence la propriété d'identité puisqu'il s'agit d'indice de Laspeyres ainsi que la propriété de circularité. La multiplication de deux indices chaînes est un indice chaîne.

$$C(p)_{4/2} \cdot C(p)_{2/0} = C(p)_{4/0}$$

$$\left[L(p)_{4/3} \cdot L(p)_{3/2}\right] \cdot \left[L(p)_{2/1} \cdot L(p)_{1/0}\right] = L(p)_{4/3} \cdot L(p)_{3/2} \cdot L(p)_{2/1} \cdot L(p)_{1/0}$$

La division d'un indice chaîne (base 100 année 0) par un indice chaîne (base 100 année 0) est un indice chaîne comme le démontre l'exemple ci-dessous. La propriété de réversibilité est respectée. Le rapport de deux indices chaînes est un indice chaîne.

$$\frac{C(p)_{5/0}}{C(p)_{3/0}} = \frac{L(p)_{5/4} \cdot L(p)_{4/3} \cdot L(p)_{3/2} \cdot L(p)_{2/1} \cdot L(p)_{1/0}}{L(p)_{3/2} \cdot L(p)_{2/1} \cdot L(p)_{1/0}} = L(p)_{5/4} \cdot L(p)_{4/3} = C(p)_{5/3}$$

$$\frac{C(p)_{5/0}}{C(p)_{4/0}} = \frac{L(p)_{5/4} \cdot L(p)_{4/3} \cdot L(p)_{3/2} \cdot L(p)_{2/1} \cdot L(p)_{1/0}}{L(p)_{4/3} \cdot L(p)_{3/2} \cdot L(p)_{2/1} \cdot L(p)_{1/0}} = L(p)_{5/4} = C(p)_{5/4}$$

Les opérations usuelles sur les indices chaînes donnent des indices chaînes. La cohérence est maintenue et les pondérations sont explicites.
Ces propriétés sont vraies pour des calculs annuels, ce n'est pas exact au plan théorique. Au plan pratique, cette difficulté reste mineure, car les évolutions infra-annuelles sont habituellement faibles.
L'indice chaîne de volume est obtenu en divisant un indice de valeur par un indice chaîne de prix.

$$\frac{V_{1/0}}{C(p)_{1/0}} = C(q)_{1/0}$$

Calcul de l'indice chaîne des prix pour l'exemple numérique.
Pour ce faire, il faut disposer des indices

$$L(p)_{1/0} = \sum_{h=1}^{n} k_0^h \cdot i(p)_{1/0}^h , \; L(p)_{2/1} = \sum_{h=1}^{n} k_1^h \cdot i(p)_{2/1}^h , \; L(p)_{3/2} = \sum_{h=1}^{n} k_2^h \cdot i(p)_{321}^h .$$

Exemple

Tableau 19. Indice $L(p)_{1/0}$.

Produits	k_0^h	$i(p)_{1/0}^h$	$L(p)_{1/0} = \sum_{h=1}^{n} k_0^h \cdot i(p)_{1/0}^h$
A	35,7	103	3 678,571429
B	32,1	104	3 342,857143
C	10,7	102	1 092,857143
D	21,4	102	2 185,714286
	100,0		10300

$$L(p)_{1/0} = \frac{10300}{100} = 103{,}0$$

Tableau 20. Indice $L(p)_{2/1}$ des prix.

Produits	k_1^h	$i(p)_{2/1}^h$	$L(p)_{2/1} = \sum_{h=1}^{n} k_1^h \cdot i(p)_{2/1}^h$
A	37,9	102	3 828,225806
B	29,8	103	2 983,870968
C	12,9	106	1 316,129032
D	19,4	103	2 012,903226
	100,0		10 141,12903

$$L(p)_{2/1} = \frac{10141{,}12903}{100} \cong 101{,}41$$

Tableau 21. Indice $L(p)_{3/2}$ des prix.

Produits	k_2^h	$i(p)_{3/2}^h$	$L(p)_{3/2} = \sum_{h=1}^{n} k_2^h \cdot i(p)_{3/2}^h$
A	37,5	104	3 900
B	29,7	105	3 117,1875
C	13,3	103	1 367,96875
D	19,5	105	2 050,78125
	100,0		10 435,9375

$$L(p)_{2/1} = \frac{10435{,}9375}{100} \cong 104{,}36$$

$$C(p)_{2/0} = L(p)_{1/0} \cdot L(p)_{2/1} \cdot L(p)_{3/2} = 1{,}03 \cdot 1{,}0141 \cdot 10{,}1436 \cdot 100 \cong 109{,}01$$

Nous disposons désormais d'une quatrième possibilité pour évaluer la hausse des prix.

L'indice de volume est simple à obtenir :

$$C(q)_{3/0} = \frac{v_{3/0}}{C(q)_{3/0}} = \frac{108}{109,1} \cdot 100 = 99,11$$

Tableau 22. Tableau récapitulatif.

Indice de valeur	Indice de prix	Indice de volume
$v_{3/0} = 108$	$L(p)_{3/0} = 109,09$	$P(q)_{3/0} = 99,03$
$v_{3/0} = 108$	$P(p)_{3/0} = 108,94$	$L(q)_{3/0} = 99,17$
$v_{3/0} = 108$	$F(p)_{3/0} = 109,01$	$F(q)_{3/0} = 99,1$
$v_{3/0} = 108$	$C(p)_{3/0} = 109,01$	$C(q)_{3/0} = 99,11$

$$P(q)_{3/0} = \frac{v_{3/0}}{L(p)_{3/0}} = \frac{108}{108,94} \cong 99,03$$

Dans l'exemple traité, les écarts entre les résultats ne sont pas très importants. Ils s'expliquent, tout d'abord, par le recours à différentes moyennes : arithmétique (Laspeyres), harmonique (Paasche) ou géométrique (Fisher). Les pondérations retenues induisent également des différences selon que l'année retenue celle de base (Laspeyres), l'année terminale (Paasche) ou les années consécutives (indices chaînes). Pour un statisticien, il serait tentant de calculer une tendance centrale de cette distribution avec toujours la même question. Laquelle retenir : le mode, la médiane, une moyenne ? Il est sans doute plus pertinent de retenir que les indices, ici de prix, sont des évaluations ayant une marge d'erreur inhérente à la réalité économique elle-même.

Les raccords d'indices

L'indice chaîne souffre aussi du phénomène de vieillissement. Le résultat obtenu est différent de celui obtenu par comparaison directe. Comme tous les indices, il est nécessaire de reconstruire périodiquement un nouvel indice. D'où le problème des raccords d'indice. Au cours du temps, des différents indices se sont succédé. Pour apprécier les évolutions sur une longue période, il est nécessaire d'utiliser des indices de bases différentes et de les raccorder en langage statistique.

Soit un indice I base 100 année 0 calculé jusqu'à l'année m, et un indice I' base 100 année g. Comment apprécier l'évolution de la grandeur entre l'année 0 et l'année t ?

($t > m$ et $m>g$)

Le schéma ci-dessous permet de visualiser le problème posé pour deux indices successifs.

Indice I			
	g	m	t
	Indice I'		

La méthodologie retenue est la suivante. Il est tout d'abord nécessaire de poser la formule d'un raccord d'indice :
$I^*_{t/0} = I'_{t/g}.\frac{I_{i/0}}{I'_{i/g}}$ avec $g \le i \le m$.

Il s'agit de la valeur qu'aurait pris l'indice I à la période t. Le rapport $cr_{ig} = \frac{I_{i/0}}{I'_{i/g}}$ est un coefficient de raccordement, il mesure l'évolution relative de l'indice I' entre la période g et la période i.

L'indice $I^*_{t/0}$ est un indice de nature indéterminée. Les formules d'indices successifs n'utilisent pas les mêmes formules de calcul ni les mêmes pondérations. Même dans le cas où les indices auraient eu recours à des formules de Laspeyres, le résultat est un indice d'une autre nature – voir les pondérations implicites. Les raccords d'indice sont la moins mauvaise modalité d'apprécier l'évolution d'une grandeur sur une période où existent plusieurs indices. Cette méthode pragmatique ne s'appuie sur aucune formalisation théorique, ce qui rend le résultat difficile à interpréter. Les raccordements n'ont pas la régularité d'un « jardin à la française », ils fournissent des ordres de grandeur extrêmement utiles. C'est une solution, d'autres sont certainement envisageables.

Exemple

Tableau 23. Exemple numérique.

	0		g	k	m		t
I	100		120	130	150		
I »			100	110	120		130

Les données disponibles permettent de calculer trois coefficients de raccordement $I'_{t/i} = \frac{I'_{t/i}}{I'_{i/g}}$

	g	k	m
$cr_{ig} = \frac{I_{i/0}}{I'_{i/g}}$	$cr_{tg} = \frac{I_{g/0}}{I'_{g/g}} = \frac{120}{100} = 1{,}2$	$cr_{kg} = \frac{I_{k/0}}{I'_{k/g}} \frac{130}{110} = 1{,}18$	$cr_{mg} = \frac{I_{m/0}}{I'_{m/g}} = \frac{150}{120} = 1{,}25$

Il apparaît trois solutions possibles :

$$I^*_{t/0} = I_{t/g} \cdot = \frac{I_{g/0}}{I'_{g/g}} = 130 \cdot \frac{120}{100} = 156 \;\; ; \;\; I^*_{t/0} = I_{t/g} \cdot \frac{I_{k/0}}{I'_{k/g}} = 130 \cdot \frac{130}{110} = 153{,}64$$

et $I^*_{t/0} = I_{t/g} \cdot \frac{I_{m/0}}{I'_{m/g}} = 130 \cdot \frac{150}{120} = 162{,}5$.

Nous disposons donc de trois estimations, l'indice appartient à l'intervalle [153,6 ; 162,5] ce qui est une première indication. En vue de disposer d'une évaluation unique, il suffit de calculer un indicateur de tendance centrale des résultats ou des coefficients. La moyenne arithmétique ou la médiane sont les candidats les plus évidents.
La médiane des évaluations est de 156. La moyenne arithmétique des évaluations est d'environ 157,38. La moyenne des coefficients de raccordement est de 1,2106 d'où l'indice : $I^*_{t/0} = 1{,}2106 * 130 \cong 157{,}38$.

Les indices de prix

Après la présentation théorique des indices, voici une présentation, synthétique, de l'indice de prix à la consommation en France.
Les indices de prix doivent leur importance au fait qu'ils sont utilisés comme références dans les négociations salariales et pour la mesure de l'évolution du pouvoir d'achat. Son utilisation dans l'estimation en volume des séries de la comptabilité nationale fait beaucoup moins polémique. L'objectif de l'indice des prix à la consommation est ainsi défini par l'INSEE :

> « L'indice des prix à la consommation est donc un instrument de mesure de l'évolution d'ensemble des prix, des biens et services figurant dans la consommation des ménages. »[1]

Cette définition toute simple soulève de nombreux problèmes, l'indice des prix à la consommation ne vise à mesurer ni le coût de la vie ni les variations du budget du consommateur moyen. De plus, l'indice des prix n'est pas un indice de dépense puisqu'il estime l'évolution des prix en considérant la consommation invariante.
Divers auteurs ont explicité les propriétés que devrait respecter un indice synthétique des prix. L'accord s'est à peu près réalisé sur les cinq conditions proposées par Lucien March[2] :

1. *Pour comprendre l'indice des prix*, Paris, INSEE.
2. Citées par Jacqueline Fourastié, *Les formules d'indices de prix*, Paris : Armand Colin, 1966.

- la simplicité, la formule retenue doit être aussi simple que possible de façon à être comprise par tous ;
- la sensibilité, quand le prix d'un produit varie l'indice global doit varier ;
- la commensurabilité, l'indice doit être indépendant des unités de mesure des quantités ;
- la proportionnalité, quand chaque indice élémentaire augmente d'un même pourcentage, l'indice synthétique doit augmenter de ce même pourcentage et
- la circularité, le rapport entre deux indices relatifs à deux dates différentes doit être indépendant de la période de base. Les changements de base doivent être possibles.

Pour comprendre les indices de prix, il est indispensable d'admettre que la hausse des prix n'existe pas en soi, en dehors d'une construction théorique.

> « En ce sens, un indice des prix à la consommation ne peut pas être seulement un instrument de mesure de l'évolution des prix des biens et services figurant dans la consommation des ménages, il est simultanément aussi la construction de la notion d'évolution des prix qu'il entend mesurer. »[3]

La construction d'un indice de prix est toujours un choix d'hypothèses et de conventions. Une critique sérieuse d'un indice de prix porte sur ces choix et non sur le fait que l'indice est une évaluation, ce qui est sa définition même.
Le premier indice de prix sérieux et rudimentaire fut établi par G. R. Carli qui voulait comparer les prix de l'huile, du vin et du grain en 1750 avec ceux de 1500[4].
Pour choisir un volume invariable, une liste de produits est déterminée avec pour chacun d'eux une quantité précise ; les prix observés sont ceux des produits de la liste. Le choix des composants et de la pondération reste à effectuer. Deux solutions sont possibles.
La première méthode, dite du budget type, se fonde sur une structure de la consommation légitime. Cette consommation normale est déterminée au sein de commissions du coût de la vie composées d'experts, de responsables politiques et de représentants des diverses catégories sociales. Cette méthode répond mieux aux exigences du calcul d'un indice du coût de la vie. En 1920, le gouvernement crée une Commission centrale d'études relatives au coût de la vie. Elle est composée de représentants des consommateurs, des employeurs, des salariés et de l'administration. Elle décide de créer

3. Jean-Paul Piriou, *L'indice des prix*, Paris : Éditions La Découverte / Maspéro, 1983, p. 16.
4. Pour une approche rapide des essais de construction des indices de prix se reporter à Jean-Louis Boursin, *Les indices de prix*, Paris : PUF, coll. « Que-sais-je ? », 1979.

60 commissions départementales chargées de calculer un indice du coût de la vie pour une famille ouvrière dans chacune de leur circonscription. Les pondérations retenues diffèrent selon les commissions et la consommation annuelle de référence varie : 48 kg de viande à Bordeaux, 300 kg à Châlons-sur-Marne, 365 kg de pain au Mans et 900 kg à Poitiers. Les indices locaux divergent tellement que la Commission centrale adopte un « budget-cadre ». L'alimentation intervient pour 60 % (9 % pour le pain), le chauffage et l'éclairage pour 5 %, le loyer pour 10 %, l'habillement pour 15 % et les dépenses diverses pour 15 %. Le défaut principal de cet indice est l'imprécision des relevés de prix.

La seconde méthode, plus empirique, consiste à observer du mieux possible les produits achetés durant la période de base par la population de référence. Cette méthode s'appuie longtemps sur les enquêtes périodiques « Budgets des familles », puis la mise en place d'un indice chaîne conduit à une mise à jour annuelle en utilisant les informations provenant de la comptabilité nationale. L'ensemble ainsi délimité constitue la base de référence permettant de définir les produits concernés et les pondérations.

En France, depuis le début du siècle du xxe siècle, sept générations d'indices de prix à la consommation se sont succédé[5]. Les progrès réalisés d'un indice au suivant ont permis d'en étendre la représentativité aussi bien en termes de population que de couverture géographique, que de composition du panier de consommation. En 1904, la Statistique générale de la France publie un indice annuel, qui devient mensuel à partir de 1911. Le modèle est la moyenne arithmétique non pondérée des rapports des prix d'un certain nombre de marchandises. Le premier indice de prix calculé en France est l'indice « 34 articles » publié de 1914 à 1949 tout d'abord en base 100 en 1914, puis en base 100 en 1938. Les ménages de référence sont les familles ouvrières de quatre personnes à Paris. Cet indice est calculé à partir d'un budget type (24 denrées alimentaires, 4 articles de chauffage et éclairage, 1 produit d'entretien). Il ne s'agit pas véritablement d'indices, les pondérations résultaient des décisions de Commissions du coût de la vie.

En 1950, un véritable indice de prix de détail est construit, celui des « *213 articles* » base 100 en 1949 ; victime de la politique de l'indice, il est abandonné en 1957. La population de référence était constituée des familles ouvrières ou employées de quatre personnes du département de la Seine. Les pondérations sont déterminées par des enquêtes sur les budgets

5. Alain Saglio, « Un nouvel indice des prix à la consommation (1990 = 100) », *Problèmes économiques*, n° 2311, 3 février 1993, repris des *Notes bleues de Bercy*, 1er novembre 1992

des familles. Les produits retenus sont l'alimentation – hors produits frais et alcools –, les produits manufacturés – sauf les biens durables, l'essence et les médicaments – et quelques services.
L'indice des « *250 articles* » qui le remplace, en base 100 juillet 1956-juin 1957, est ensuite calculé de 1957 à 1962. La population de référence se modifie. Ce sont toujours les ménages d'ouvriers ou d'employés mais de plus de deux personnes habitant la Seine ou 17 capitales régionales. Les produits repris dans l'indice et leurs pondérations sont obtenus par enquêtes sur les budgets des familles.
À partir de 1963, il est remplacé par un indice national des « *259 articles* », base 100 en 1962, qui sera calculé jusqu'en 1970. La population prise en compte est identique, les lieux d'habitation sont, par contre, étendus à toutes les agglomérations de plus de 2 000 habitants. Le champ des produits retenus s'étend à l'alimentation (hors alcool), à tous les produits manufacturés et à 60 % des services.
Il laisse la place, en 1971, à l'indice des « *295 postes de dépenses* » base 100 en 1970, puis base 100 en 1980, calculé de 1971 à 1992. Il devient indice des « *296 postes* » en 1987 avec introduction des services bancaires. Il estime l'évolution des prix pour les dépenses des ménages dont le chef est ouvrier ou employé, pour toutes les agglomérations de plus de 2 000 habitants. Les pondérations sont mises à jour annuellement selon la procédure de l'indice chaîne. Il comprend toute l'alimentation, tous les biens manufacturés et 80 % des services. Cet indice a été robuste, puisqu'en 22 ans de service il n'a subi que des aménagements mineurs : base 100 en 1980, introduction du service des banques, prise en compte de la vente par correspondance.
Cet indice est remplacé depuis 1993 par l'indice des « *265 postes* » en base 100 en 1990 qui est un indice « tous ménages ». L'abandon de l'indice des « 295 postes » s'explique par plusieurs raisons. Principalement, l'indice avait vieilli et les dépenses fixes ne correspondaient plus à la structure des dépenses. Des facteurs externes interviennent dans l'obsolescence de l'indice : l'adaptation nécessaire aux conventions de la comptabilité nationale, la mise en cause par les organisations syndicales, les besoins de comparaisons internationales, l'utilisation de nouvelles techniques plus efficaces (informatique), etc.
Le changement de la population de référence du nouvel indice s'explique par le fait que les ménages dont le chef est ouvrier ou employé ne représentent plus que 25 % des ménages au recensement de 1990. Ce faible taux de couverture affaiblissait son utilité dans les comparaisons internationales, car beaucoup – Allemagne, Pays scandinaves – calculent des indices pour l'ensemble de la population. De plus, cet indice avait vieilli en ce sens qu'il ignorait certains services – assurances, transports aériens… De plus, il s'articulait mal avec les comptes nationaux suite aux changements de base.

L'utilisation de la comptabilité nationale conduit à élargir le champ de la population de référence de l'indice. Sa conception rendait difficiles les comparaisons internationales et au sein de la CEE. Sa construction répond aussi à un rapprochement vers une harmonisation des indices de prix de la CEE puis de l'Union européenne.

L'outil d'observations statistiques doit s'adapter périodiquement pour rester pertinent. L'indice des prix à la consommation est modifié à compter du premier janvier 1993. Les modifications répondent à une meilleure description de l'évolution des prix à la consommation ainsi qu'à la recherche d'une plus grande lisibilité des comparaisons internationales. La rénovation de l'indice permet un rapprochement avec les indices des prix à la consommation des principaux pays européens ; c'est le cas en ce qui concerne la population de référence. Les principales caractéristiques sont décrites ci-dessous. L'indice actuel a fait l'objet d'une révision en base 1998.

L'indice des prix à la consommation (IPC) base 1998, décomposé en 161 groupes, 86 regroupements et 12 fonctions de consommation, constitue la septième génération d'indice.

La consommation est définie au sens de la comptabilité nationale. Les achats et constructions d'immeubles, y compris les logements – et les dépenses assimilées comme les dépenses de gros entretiens des logements et des immeubles –, ainsi que les achats de valeurs mobilières ne sont pas pris en compte car ils sont considérés comme des investissements. Ce qui exclut la prise en compte de dépenses liées, comme le coût du crédit au logement.

Les biens durables comme les automobiles, les meubles, ne sont pris en compte par l'indice que s'ils sont à l'état neuf ; les achats d'occasions sont analysés comme des transferts de ressources au sein des ménages. Le nouvel indice prend en compte les transports aériens, maritimes et côtiers, les locations de voiture, les transports par ambulance, les services vétérinaires et les services funéraires. L'indice couvre 95 % des dépenses de consommation des résidents et des touristes ; en sont exclus les dépenses d'hospitalisation privée, les services d'assurance vie et les jeux de hasard, les opérations d'épargne et les cotisations sociales, les paiements partiels de services non marchands comme les crèches ou les contributions résiduelles de santé. En sont exclues les assurances vie car elles répondent certes à la couverture d'un risque (incluses dans la consommation) mais elles correspondent également à des placements financiers (exclus de la consommation). Les services domestiques sont toujours exclus de l'indice, en particulier car les dépenses afférentes sont mal connues. Les impôts directs et le coût du crédit à la consommation, les frais d'hospitalisation sont exclus du calcul de l'indice. Par contre, les fluctuations de la TVA interviennent dans le calcul de l'indice. Puisque, les prix sont relevés, TVA comprise, la modification du taux de TVA joue sur l'indice des prix.

Les relevés de prix sont réalisés auprès de points de vente pour les biens et les services et les tarifs sont fournis par les organismes producteurs (électricité, télécommunications, transport ferroviaire ou aérien, vente à distance…). La collecte des informations est réalisée chaque mois et tous les quinze jours pour les produits frais. Elle est réalisée dans 106 agglomérations de plus de 2000 habitants dans les différentes formes de vente, y compris internet. La liste des produits de l'échantillon est confidentielle, ce sont environ 130 000 séries produits précis qui sont suivies permettant 160 000 relevés mensuels auxquels s'ajoutent 40 000 tarifs. L'échantillon est remis à jour annuellement pour tenir compte de la disparition de produits, de l'apparition de nouveaux produits et des changements de structure de la consommation. L'IPC est un indice chaîne de Laspeyres, dont les pondérations sont modifiées chaque année en fonction de l'évolution des consommations de la disparition et de l'apparition des produits (biens ou services). Régulièrement, des enquêtes approfondies permettent de le rénover, il est également modifié en relation avec les indices de prix utilisés par l'Union européenne.

Tableau 24. Structure des indices prix (base 100 : année 1998).

Regroupements	Pondérations 2014
a) Ensemble des ménages - France	
ENSEMBLE	10 000
ALIMENTATION	1 653
Produits frais	210
Alimentation hors produits frais	1 443
TABAC	204
PRODUITS MANUFACTURÉS	2 653
Habillement et chaussures	438
Produits de santé	434
Autres produits manufacturés	1 781
ÉNERGIE	850
dont Produits pétroliers	475
SERVICES	4 640
Loyers, eau et enlèvement des ordures ménagères	748
Services de santé	552
Transports et communications	503
Autres services	2 837
ENSEMBLE HORS LOYERS ET HORS TABAC	9 202
ENSEMBLE HORS TABAC	9 796
b) Ménages urbains dont le chef est ouvrier ou employé	
ENSEMBLE HORS TABAC	9 707
ENSEMBLE	10 000
c) Ménages du 1er quintile de la distribution des niveaux de vie	
ENSEMBLE HORS TABAC	9 679

Source : INSEE – indices des prix à la consommation

La présentation des indices est incontournable dans le champ de l'économie. Les indices, particulièrement celui des prix à la consommation, font l'objet de nombreuses controverses en raison de l'importance des enjeux politiques et sociaux associés. Nous avons montré que du point de vue théorique, il était possible de calculer sur une même distribution de prix élémentaires plusieurs indices tous absolument justes qui peuvent donner des résultats différents. Les écarts constatés entre les indices théoriques sont plus faibles qu'entre les dispositifs mis en œuvre pour les calculer en pratique. Les résultats sont toujours des ordres de grandeur ce qui est déjà très important. Dans des situations économiques stables, avec des hausses de prix fables et sans bouleversements économiques majeurs, les différents indices donnent des résultats très similaires, les différences restent inférieures aux erreurs de calcul.

Conclusion

Cet ouvrage de statistique a présenté les concepts et les méthodes employées pour traiter les données. Il a aussi permis d'appréhender les techniques utilisées pour obtenir des caractéristiques synthétiques pertinentes pour mesurer ou apprécier la situation d'une économie et son évolution. La statistique descriptive ignore les éléments aléatoires et s'en tient à une démarche déterministe considérant que l'échantillon étudié est représentatif, ce qui est souvent le cas pour nombre de données économiques. Par contre, si les enquêtes se multiplient, les résultats obtenus seront également plus nombreux et très probablement différents. L'ensemble forme alors une distribution statistique pour laquelle il sera possible de calculer une valeur centrale. En statistique descriptive, cette valeur centrale calculée, associée éventuellement, ce qui est exceptionnel pour les données publiées, à une caractéristique de dispersion, sera considérée comme *exacte*. Il est cependant impossible d'affirmer que cette valeur est *vraie*. Les méthodes de la statistique inférentielle associent à cette valeur une probabilité et une estimation des risques d'erreur pris en acceptant cette valeur. D'autres techniques ont pour objet de préciser les conditions de généralisation à une population des résultats issus d'un ou de plusieurs échantillons d'où le recours aux lois de probabilités et à l'analyse des données en masse, c'est le domaine, la statistique fréquentielle. Ces approches sont d'autant plus indispensables pour compléter l'information fournie par la statistique descriptive que les données économiques sont généralement frappées d'imprécision tenant à la nature des données et des réalités des populations statistiques étudiées.

Les outils de la statistique descriptive permettent de produire des indicateurs propres aux divers champs et domaines de l'analyse économique et de l'économie politique. Le renouveau des réflexions sur les statistiques et les indicateurs s'explique pour partie par les considérations économiques, politiques, sociales et éthiques en relation avec l'émergence du paradigme du développement durable. En effet, pour évaluer les transformations liées au processus connu sous le nom de développement durable ou développement

soutenable, des indicateurs sont nécessaires. Le paradigme de la croissance s'est longtemps appuyé sur le calcul d'un indicateur unique le produit intérieur brut (PIB) souvent confondu avec une mesure du développement en désaccord avec les théoriciens de l'économie qui distinguaient clairement les deux notions. Le concept de développement durable, avec toutes les ambiguïtés de sa définition, retient une approche large du développement dépassant la seule croissance économique. Les indicateurs doivent permettre de mesurer le développement dans trois dimensions économique, sociale et environnementale. Des conférences internationales, des commissions, des colloques universitaires sont régulièrement tenus pour valider ou proposer des indicateurs pertinents et fiables. Ces débats sont loin d'avoir abouti à des solutions consensuelles. Les propositions multiples se doivent de déboucher sur des calculs et sur des résultats compréhensibles pour la majorité des citoyens. Elles utilisent très largement les concepts et les outils de la statistique descriptive que ce soit en termes de tendance centrale, soit de dispersions. La compréhension des méthodologies développées dans cet ouvrage devrait permettre de garder l'esprit critique dans l'interprétation des solutions proposées. La partie du chapitre 5 dédié aux indices de prix a montré que les calculs répondaient à des choix et des contraintes et que plusieurs solutions satisfaisantes pouvaient coexister. Ce simple exemple suffit pour inciter à toujours s'interroger sur les conditions d'obtention des données, sur les méthodes utilisées pour obtenir des résultats quantifiés et sur leurs domaines de validité en fonction des objectifs poursuivis. La construction d'indicateur est un processus social complexe, les outils statistiques ne peuvent s'y substituer.

En économie et en sciences sociales plus généralement, la mesure ne consiste pas simplement à quantifier une réalité préalable, comme ce peut être le cas pour certaines mesures physiques, elle définit également ce qui est mesuré. L'objet de la mesure doit être nécessairement délimité au préalable et ainsi lui donner une réalité sociale ou économique. L'exemple de la mesure du chômage permet d'illustrer cette démarche. Le concept de chômage n'existe pas préalablement à son traitement social[1], alors que la situation d'être sans emploi et d'en chercher un existant bien antérieurement. Le chômage est une construction sociale et institutionnelle. Les personnes respectant les critères de définition obtiennent des droits à condition d'accepter les vérifications

1. Salais R., Reynaud B. et N. Baverez. (1999), *L'invention du chômage : histoire et transformations d'une catégorie en France des années 1890 aux années 1980*, Paris, Presses universitaires de France, coll « Quadrige ».

imposées pour la bonne gestion des flux. La production d'indicateurs permet de disposer d'un langage commun pour permettre le dialogue qui passe par des concepts largement acceptés et par des données fiables et robustes. La demande pour des données statistiques fiables et facilement accessibles est croissante en raison des facilités de la diffusion permise par les technologies de l'information. Les débats français sur l'indice des prix à la consommation et la mesure du chômage illustrent parfaitement cette situation. La présentation et la diffusion des résultats sont l'objet de réflexions de la part des producteurs et des utilisateurs des indicateurs. L'INSEE au moment de sa création devait fournir des données aux administrations publiques et plus particulièrement au ministère des Finances, auquel il est encore rattaché[2]. Les évolutions et les demandes vers plus de transparence et de démocratie ont conduit à étendre les champs couverts par les indicateurs et à rendre l'information plus accessible au public grâce à un site Internet. Le défi pour aujourd'hui est de trouver les moyens de diffuser les informations statistiques de façon à ce qu'elles soient comprises, aussi bien par les spécialistes que par les citoyens. Nous le voyons, les débats sur les indicateurs sont plus heuristiques que pratiques. Ces interrogations soulignent néanmoins le caractère contingent tant sur le plan historique que géographique des indicateurs[3].

2. La direction de l'INSEE affiche sa préférence pour une autonomisation de l'Institut. « L'indépendance de la statistique, c'est l'indépendance de l'institution statistique ». Cette possibilité a été reprise au plan politique.

3. Cassiers I. et G. Thiry (2009), « Au-delà du PIB : réconcilier ce qui compte et ce que l'on compte », *Regards économiques*, n° 75.

Table des matières